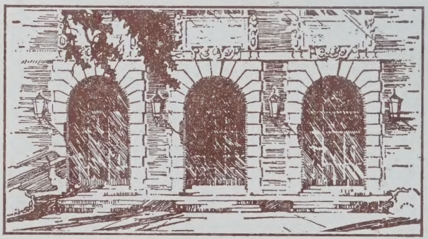
LIBRARY OF THE
UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

621.31921
Al54p
1976



ENGINEERING

MAY 4 1977

The person charging this material is responsible for its return to the library from which it was withdrawn on or before the **Latest Date** stamped below.

Theft, mutilation, and underlining of books are reasons for disciplinary action and may result in dismissal from the University.

UNIVERSITY OF ILLINOIS LIBRARY AT URBANA-CHAMPAIGN

ENGINEERING

JUN 24 1977

MAY 31 REC'D

SEP 03 1988

PHOTO REPRODUCTION

NOV 10 REC'D

AUG. 21 REC'D

MAR 6 1979

FEB 28 Rec'd

PHOTO REPRODUCTION

DEC 17 REC'D

DEC 01 1982

NOV 03 REC'D

OCT 24 1987

SEP 26 REC'D

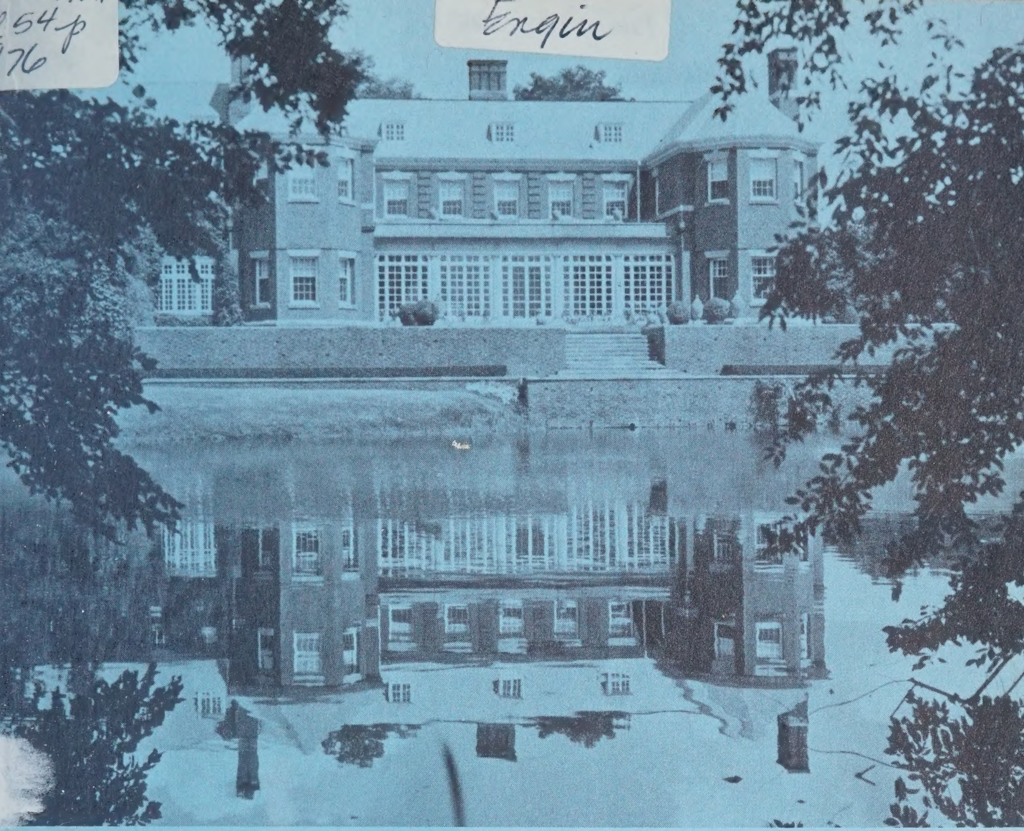
L161—O-1096

Digitized by the Internet Archive
in 2026 with funding from
University of Illinois Urbana-Champaign

https://archive.org/details/proceedings1419alle_0

54p
76

Engin



Proceedings

FOURTEENTH ANNUAL ALLERTON CONFERENCE ON CIRCUIT AND SYSTEM THEORY

September 29, 30, October 1, 1976

Allerton House, Monticello, Illinois
Sponsored by the
Department of Electrical Engineering and the
Coordinated Science Laboratory of the
University of Illinois at Urbana-Champaign

THE LIBRARY OF THE
FEB 21 1977
UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

ENGINEERING LIBRARY
UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS

2'0

PROCEEDINGS
FOURTEENTH ANNUAL ALLERTON CONFERENCE
ON CIRCUIT AND SYSTEM THEORY

T. N. Trick
M. E. VanValkenburg
Conference Co-Chairmen

Conference held

September 29-October 1, 1976
Allerton House
Monticello, Illinois

Sponsored by

The Department of Electrical Engineering
and
The Coordinated Science Laboratory
of the
UNIVERSITY OF ILLINOIS
at
Urbana-Champaign

TABLE OF CONTENTS

Foreword iii

I-a: RANDOM PROCESSES
Chairman and Organizer: Donald L. Snyder, Washington University

AN INTRODUCTION TO DISCRETE TIME MARTINGALES AND FILTERING PROBLEMS 1
J. H. Van Schuppen, Washington University

OPTIMAL, CAUSAL, SIMULTANEOUS DETECTION AND ESTIMATION OF RANDOM SIGNAL FIELDS IN GAUSSIAN NOISE FIELD 10
T. T. Kadota, Bell Telephone Laboratories, Murray Hill, New Jersey

THE RECONSTRUCTION OF RANDOM SIGNALS FROM IRREGULARLY SPACED SAMPLES 19
Frederick J. Beutler, The University of Michigan

ANALYSIS OF PSEUDO-RANDOM NUMBER SEQUENCES GENERATED BY A METHOD OF RADER, RABINER, AND SCHAFER 27
Ting-Pien Fang and Charles E. Molnar, Washington University

ESTIMATION AND CONTROL PERFORMANCE FOR SPACE-TIME POINT-PROCESS OBSERVATIONS 38
Ian B. Rhodes and Donald L. Snyder, Washington University

I-b: INFINITE DIMENSIONAL AND TIME DELAY SYSTEMS - I
Chairman: A. Manitius, Université de Montréal
Co-Organizers: A. Manitius and M. C. Delfour, Université de Montréal

TOWARD A GENERAL THEORY OF CONTROL CANONICAL STRUCTURE 52
David L. Russell, University of Wisconsin-Madison

DIFFERENTIAL-DIFFERENCE EQUATIONS IN A HILBERT SPACE . 62
Richard Datko, Georgetown University

QUALITATIVE ANALYSIS OF INTERCONNECTED DYNAMICAL SYSTEMS DESCRIBED ON BANACH SPACES: WELL POSEDNESS, BOUNDEDNESS AND INSTABILITY 70
Anthony N. Michel and Richard K. Miller, Iowa State University

ON THE RELATIONSHIP BETWEEN FIRST AND SECOND ORDER CONTROLLABLE SYSTEMS IN BANACH SPACES 80
Roberto Triggiani, Iowa State University

OPERATOR THEORY AND BROADBAND MATCHING 91
J. William Helton, University of California, San Diego, La Jolla, California

I-c: STABILITY AND STABILIZATION
Chairman: J. Medanić, University of Illinois

VULNERABILITY OF DYNAMIC SYSTEMS 99
D. D. Šiljak, University of Santa Clara

ON STRUCTURAL STABILITY OF MULTIVARIABLE SYSTEMS . . . 113
E. Freund, Institut für Informationsverarbeitung in Technik und Biologie der Fraunhofer-Gesellschaft, West Germany

ON A RESULT BY HESTENES AND ITS APPLICATION IN BILINEAR SYSTEMS STABILIZATION 122
Tudor Ionescu, Intermetrics, Inc., Cambridge, Mass.

POLE PLACEMENT DESIGN OF MULTIPLE CONTROLLER SYSTEMS VIA WEAK AND STRONG CONTROLLABILITY 132
Joe H. Chow, University of Illinois

A CONDITION FOR DECENTRALIZATION IN MODEL REFERENCE ADAPTIVE SYSTEMS 140
Malur K. Sundareshan, Concordia University, Montreal, Canada

INPUT-OUTPUT STABILITY OF INTERCONNECTED SYSTEMS DESCRIBED BY NONLINEAR STOCHASTIC DIFFERENTIAL EQUATIONS 147
R. L. Gutmann and A. N. Michel, Iowa State University

I-d: FAULT TOLERANT SYSTEMS
Chairman: Gernot Metze, University of Illinois

AN INTEGRATED CIRCUIT FAULT MODEL FOR DIGITAL SYSTEMS . 156
Mark Ketelsen, Northwestern University

SELF-CHECKING DESIGN USING COMPLETE SETS OF ALTERNATING PRIMITIVES 166
Dennis A. Reynolds, Sandia Laboratories, Albuquerque, New Mexico

A FUNCTIONAL FORM APPROACH TO FAULT INTERRELATIONSHIPS IN TREE NETWORKS 178
Vinod K. Agarwal and Gerald M. Masson, The Johns Hopkins University

SOME REDUNDANCY AND COMPLETE MASKING OBSERVATIONS . . 188
Vinod K. Agarwal and Gerald M. Masson, The Johns Hopkins University

A LOW COST TOTALLY SELF-CHECKING CHECKER FOR RESIDUE 3 AND 3N CODES 189
Jean Dussault and Gernot Metze, University of Illinois

DESIGN OF ASYNCHRONOUS CIRCUITS -- SOME PROBLEMS . . 191
Jon G. Kuhl and S. M. Reddy, The University of Iowa

I-e: DIGITAL FILTERS
Chairman: T. N. Trick, University of Illinois

PROPERTIES OF THE COMPLEX DIGITAL INVERSE FILTER AS APPLIED TO NARROWBAND INTERFERENCE REJECTION 201
Frank W. Symons, Jr., The Pennsylvania State University

FOURIER SERIES REPRESENTATION FOR POLYNOMIALS WITH APPLICATION TO NONLINEAR DIGITAL FILTERING 211
Neal C. Gallagher, Jr., Purdue University, John W. Allen, Case Western Reserve University, and Gary L. Wise, University of Texas at Austin

STEADY-STATE, ROUNDOFF-ERROR ANALYSIS IN DIGITAL FILTERS REALIZED WITH FIXED-POINT ARITHMETIC 219
A. Sripad and D. L. Snyder, Washington University

NECESSARY AND SUFFICIENT CONDITIONS FOR QUANTIZATION-ERROR STATISTICS TO BE UNIFORM AND WHITE 221
A. B. Sripad and D. L. Snyder, Washington University

DIGITAL OSCILLATORS WITH INHERENT NONLINEARITY . . . 222
Hanoch Ur and Gershon Leshem, Tel Aviv University, Israel

SOME LOWPASS NARROW-BAND LOW-NOISE FIXED-POINT RECURSIVE DIGITAL FILTER STRUCTURES 233
D. J. Mayer, Texas Instruments, Inc., Dallas, Texas and T. N. Trick, University of Illinois

NEW RESULTS ON THE BLOCK IMPLEMENTATION OF RECURSIVE DIGITAL FILTERS 243
S. K. Mitra and R. Gnanasekaran, University of California, Davis, California

II-a: POWER SYSTEMS

Chairman and Organizer: A. M. Sasson,
American Electric Power

STOCHASTIC POWER FLOW STUDY METHODS 253
G. T. Heydt and P. W. Sauer, Purdue University

NONLINEAR RESISTIVE CIRCUIT MODELS FOR POWER SYSTEM STEADY-STATE ANALYSIS 261
Felix F. Wu, University of California, Berkeley and R. L. Sullivan, Energy Research and Development Administration, Washington, D.C.

THE POTENTIAL FOR APPLICATION OF LINEAR AND QUADRATIC PROGRAMS IN SOLVING ECONOMIC DISPATCH AND EXPANSION PLANNING PROBLEMS 269
Gyaneshwar Dayal, Harza Engineering Company, Chicago, Illinois

A COMPARATIVE STUDY OF POWER NETWORK ALGORITHMS . . . 279
Fernando L. Alvarado, The University of Wisconsin

SECURITY, CONTROL, AND ECONOMIC FUNCTIONS IN POWER SYSTEM CONTROL CENTERS 289
Mark Enns, The University of Michigan, Ann Arbor

THE Z-MATRIX IN POWER SYSTEM NETWORK ANALYSIS 299
H. E. Brown, North Carolina State University

II-b: INFINITE DIMENSIONAL AND TIME DELAY SYSTEMS - II

Chairman: M. C. Delfour, Université de Montréal
Co-Organizers: A. Manitius and M. C. Delfour,
Université de Montréal

THE METHOD OF STEPS IN THE THEORY OF CONTROL OF TIME DELAY PROCESSES 308
E. B. Lee, University of Minnesota

STATE AND INPUT FEEDBACK IN SYSTEMS CONTAINING TIME DELAYS 310
Edward W. Kamen, Georgia Institute of Technology

FUNCTION SPACE CONTROLLABILITY OF RETARDED SYSTEMS: SOME NEW ALGEBRAIC CONDITIONS 320
A. Manitius, Université de Montréal, Canada

INDIRECT DELAY-FEEDBACK CONTROL OF LINEAR SYSTEMS . . 328
B. Asner, Jr., University of Dallas, Irving, Texas and A. Halanay, University of Bucharest, Romania

II-c: INFORMATION, GAMES AND DECENTRALIZED CONTROL
Chairman: I. B. Rhodes, Washington University
Co-Organizers: I. B. Rhodes, Washington University and J. B. Cruz, Jr., University of Illinois

COMMON AND PRIVATE INFORMATION BETWEEN TWO PERSONS . . 338
Y. C. Ho, Harvard University

DECENTRALIZED ROBUST CONTROL OF UNKNOWN SYSTEMS USING TUNING REGULATORS 339
E. J. Davison, University of Toronto, Canada

ON STOCHASTIC GAMES WITH NONCOMPACT STRATEGY SPACES AND LOSS FUNCTIONS THAT EXHIBIT A PRODUCT STRUCTURE: WITH APPLICATION TO LEG STOCHASTIC DIFFERENTIAL GAMES. 349
Mary Ellen Bock, Purdue University and
Max Mintz, University of Pennsylvania

OPTIMAL CONTROL OF TWO-LEVEL DYNAMIC SYSTEMS 359
M. Simaan, University of Pittsburgh

ON THE MULTILEVEL STACKELBERG STRATEGIES IN LINEAR QUADRATIC SYSTEMS 369
J. Medanić, University of Illinois and
D. Radojevic, "Mihailo Pupin" Institute,
Belgrade, Yugoslavia

II-d: LINEAR SYSTEMS
Chairman: William R. Perkins, University of Illinois

A NEW APPROACH TO OUTPUT FEEDBACK STABILIZATION . . . 380
Adly T. Fam, University of California, Irvine, California and James S. Meditch, University of Washington

AUTOMATIC GENERATION CONTROL: A DECENTRALIZED APPROACH 389
M. S. Ćalović, Serbian Electric Power Board, Belgrade, Yugoslavia, M. V. Đorović, City of Belgrade Headquarters, Belgade, Yugoslavia, and D. D. Šiljak, University of Santa Clara

AN APPROACH TO POLE ASSIGNMENT BY EXTERIOR ALGEBRA . . 399
V. Seshadri and M. Sain, University of Notre Dame

SAMPLED-DATA OBSERVABILITY OF CONTINUOUS-TIME SYSTEMS 408
Robert A. Schlueter, Michigan State University and Richard K. T. Ma, IBM Thomas J. Watson Research Center, Yorktown Heights, New York

LINEAR SYSTEM ESTIMATES VIA INTERMEDIATE DOMAIN PROCESSING 418
Sorin Cohn-Sfetcu, Bell-Northern Research, Ottawa, Canada

A NORMALIZED RESIDUAL CRITERION FOR THE DETERMINATION OF STRUCTURE OF LINEAR SYSTEMS 423
L. C. Suen and R. Liu, University of Notre Dame

II-e: RANDOM PROCESSES AND APPLICATIONS
Chairman: T. T. Kadota, Bell Laboratories

LOCAL DETECTION OF STOCHASTIC SIGNALS IN ADDITIVE NOISE 432
H. V. Poor and J. B. Thomas, Princeton University

PROPERTIES OF CERTAIN CLASSES OF CAUSAL, NONLINEAR, TIME-VARYING STOCHASTIC SYSTEMS 441
William L. Root, The University of Michigan, Ann Arbor

ON THE REPRESENTATION OF NONLINEAR SYSTEMS WITH GAUSSIAN INPUTS 451
Steel T. Huang, University of Cincinnati and Stamatis Cambanis, University of North Carolina

A REPRESENTATION FOR SPHERICALLY INVARIANT RANDOM PROCESSES 460
G. L. Wise, University of Texas at Austin and N. C. Gallagher, Jr., Purdue University

ERROR-PROBABILITY BOUNDS FOR CONTINUOUS-TIME DECISION PROBLEMS 470
J. L. Hibey, D. L. Snyder, and J. H. Van Schuppen, Washington University

THE ZERO CROSSING INTERVAL STATISTICS OF THE SMOOTHED RANDOM TELEGRAPH SIGNAL 471
John T. Rickard, Orincon Corporation, La Jolla, California

POWER MEAN FREQUENCY ESTIMATION FOR A RANDOM PROCESS GATED AT A SUB-NYQUIST RATE 481
Jayvadan N. Patel and Lester C. Peach, University of Chicago and Illinois Institute of Technology

III-a: MODELING AND IDENTIFICATION
Chairman and Organizer: William L. Root
University of Michigan

SIMULATION OF ABNORMAL CELL BEHAVIOUR 482
E. J. Davison, University of Toronto, Canada

COMPARISON OF DIFFERENT CLASSES OF DYNAMIC MODELS . . 491
R. L. Kashyap, Purdue University

LINEAR PREDICTIVE CODING WITH ADDITIVE NOISE FOR APPLICATION TO SPEECH DIGITIZATION 500
James L. Melsa and James D. Tomcik, University of Notre Dame

IDENTIFICATION PROBLEMS IN ANTICOAGULANT DRUG THERAPY 509
William F. Powers and Peter H. Abbrecht, University of Michigan, Ann Arbor

RECURSIVE STATE OBSERVATION AND PARAMETER IDENTIFICATION USING HYPERSTABILITY THEORY 517
D. M. Boroson and S. C. Schwartz, Princeton University

III-b: INFINITE DIMENSIONAL AND TIME DELAY SYSTEMS - III
Chairman: R. Triggiani, Iowa State University
Co-Organizers: A. Manitius and M. C. Delfour,
Université de Montréal

TIME OPTIMAL CONTROL OF HEREDITARY SYSTEMS 528
M. Q. Jacobs and W. C. Pickel, University of Missouri

FILTERING OF LINEAR STOCHASTIC HEREDITARY DIFFERENTIAL SYSTEMS: NUMERICAL SOLUTION 535
Michel Delfour, Université de Montréal, Canada

THE LINEAR QUADRATIC GAUSSIAN PROBLEM FOR SYSTEMS WITH DELAYS IN THE STATE, CONTROL, AND OBSERVATIONS . 545
R. H. Kwong, McGill University and Université de Montréal, Canada

REGULATOR THEORY FOR EVOLUTION SYSTEMS 550
K. P. M. Bhat, University of Toronto, Canada

III-c: ACTIVE FILTERS

Chairman: M. E. VanValkenburg, University of Illinois

OPTIMUM NOTCH FILTERING FOR PULSED COMMUNICATION SIGNALS 558
Richard W. Koralek, San Jose State University and Armin C. W. Reese, E S L, Inc., Sunnyvale, California

NEW BIQUADS USING TWO OPERATIONAL AMPLIFIERS 559
Keiji Suzuki, Nippon Electric Company, Ltd., Kawasaki, Japan

ACTIVE RC MULTIPORT SYNTHESIS OF A TRANSFER ADMITTANCE MATRIX WITH A MINIMUM NUMBER OF CAPACITORS 569
Khalil Ahmad Khan and Afaq Ahmad, Aligarh Muslim University, India

COMPUTER-AIDED TOLERANCE AND CORRELATION COEFFICIENT ASSIGNMENT: A FIRST ATTEMPT 578
Arthur R. Thorbjornsen and M. Malek Daaboul, The University of Toledo

STATISTICAL MULTIPARAMETER SENSITIVITY MEASURES FOR TRANSFER, GAIN AND PHASE FUNCTIONS 587
Cevdet Acar, Wayne State University

III-d: DIGITAL SYSTEMS - I

Chairman: J. A. Abraham, University of Illinois

ON LIVENESS, DEADLOCK, AND REACHABILITY OF E-NETS . . 597
Tadao Murata and Tasvir Shah, University of Illinois at Chicago Circle

CHARACTERISTICS OF SOME AUGMENTED PETRI NETS 606
J. E. Smith, University of Wisconsin, Madison and T. Mudge, University of Illinois

REDUCTION OF PETRI NETS 616
Larry L. Kinney, University of Minnesota and Yih-Wu Han, Honeywell, Inc., Minneapolis, Minnesota

THE CONTROL PROBLEM FOR SEQUENTIAL MACHINES 626
David W. Fraley, Battelle Northwest Laboratories, Richland, WA. and Joseph J. Talavage, Purdue University

III-e: OPTIMIZATION AND OPTIMAL CONTROL
Chairman: D. Bertsekas, University of Illinois

A NEW THEORETICAL FRAMEWORK FOR FINITE HORIZON STOCHASTIC CONTROL 636
Steve Shreve and Dimitri Bertsekas, University of Illinois

A REACHABILITY THEORY FOR A CLASS OF UNCERTAIN DYNAMICAL SYSTEMS 644
B. Ross Barmish, Yale University

COMPUTATIONAL METHODS TO OBTAIN THE OPTIMAL JET ENGINE CONTROL 652
R. J. Basso and R. J. Leake, University of Notre Dame

OPTIMAL CONTROL OF DELAYED ACTION CONTROL SYSTEM IN FUNCTION SPACE 662
Z. Bien and D. H. Chyung, The University of Iowa

IV-a: CIRCUIT ANALYSIS TECHNIQUES
Chairman: Nelson Wax, University of Illinois

A NEW ANALYTICAL REPRESENTATION FOR SECTION-WISE PIECEWISE-LINEAR FUNCTIONS AND ITS APPLICATIONS . . . 667
Leon O. Chua, University of California, Berkeley and Sung Mo Kang, Rutgers University

SIGNAL PROCESSING VIA DISTRIBUTED FEED NETWORKS . . . 677
D. Hazony, Case Western Reserve University

A NEWTON'S ALGORITHM FOR SOLVING SYSTEMS OF NON-LINEAR EQUATIONS AND INEQUALITIES 683
H. Muaki, Washington University

IV-b: CODING THEORY
Chairman: John B. Thomas, Princeton University

PRIMITIVE CODING THEORY FOR CERTAIN CLASSES OF DISCRETE CHANNELS 684
William L. Root, The University of Michigan

ON TRUNCATED VITERBI DECODING 694
Daniel J. Costello, Jr. and Farhad Hemmati, Illinois Institute of Technology

REDUCED-STATE VITERBI DECODING OF CONVOLUTIONAL CODES 695
Jean Conan and David Haccoun, Ecole Polytechnique, Montreal, Canada

BINARY MOST SIGNIFICANT BIT CODES 704
Larry A. Dunning and W. E. Robbins, North Carolina State University

ON UNIQUELY DECIPHERABLE CODES WITH GIVEN COMPOSITIONS 712
S. C. Ntafos and S. L. Hakimi, Northwestern University

EFFICIENT INFORMATION FEEDBACK FOR BINARY ENCODED SIGNALS 722
John J. Metzner, Wayne State University

A THEOREM ON CYCLIC CODES WITH APPLICATIONS 731
Gerald E. Seguin, Royal Military College of Canada, Kingston, Canada and Vijay K. Bhargava, Concordia University, Montreal, Canada

PERFORMANCE ANALYSIS OF BINARY SEQUENTIAL COMMUNICATION SYSTEMS EMPLOYING HARD DECISION STRATEGY 732
Enzo Dalle Mese, University of Pisa, Italy and Dino Giuli, University of Florence, Italy

IV-c: SIGNAL PROCESSING AND PATTERN RECOGNITION
Chairman: S. C. Schwartz, Princeton University

OPTIMUM LINEAR IMAGE RESTORATION OF REALISTIC DATA BASED ON THE TWO-DIMENSIONAL ARMA MODEL 740
Hlung-Wen Penm and Morton Kanefsky, University of Pittsburgh

ORTHOGONAL TRANSFORM COMPARISON USING ERTS DATA . . . 750
George W. Zobrist, The University of Toledo

NOVEL USES OF QUADRATIC SURFACES FOR MEDICAL DIAGNOSIS 753
N. DeClaris, O. Ijaola, and R. W. Newcomb, University of Maryland

A SEQUENTIAL PROCEDURE FOR COSTLY DATA CLUSTERS . . . 763
P. Papantoni-Kazakos and R. Li, Rice University

BOOLEAN FLOW THEORY AND ITS APPLICATIONS TO CLUSTER ANALYSIS, FUZZY LOGICS AND PARTICLE TRANSMISSION . . 772
Kenji Onaga, Osaka University, Japan and Wataru Mayeda, University of Illinois

A NUMERICAL METHOD FOR THE COMPUTATION OF PROBABILITY OF ERROR IN A STRUCTURED DETECTOR OF AN OPTICAL COMMUNICATION SYSTEM 781
P. R. Kumar and Lester F. Eastwood, Jr., Washington University

OPTIMAL CHOICE OF THE KERNEL FUNCTION FOR KERNEL TYPE DENSITY ESTIMATORS 787
Dimitri Kazakos, State University of New York at Buffalo

RANGE OF VALIDITY OF THE SAMPLING THEOREM 792
Balaji Swaminathan and R. J. Leake, University of Notre Dame

IV-d: SINGULAR PERTURBATIONS IN CONTROL SYSTEMS
Chairman and Organizer: Petar V. Kokotović
University of Illinois

SINGULAR PERTURBATION ANALYSIS APPROACH FOR SYSTEMS WITH HIGHLY COUPLED DYNAMICS 793
Anthony J. Calise, Dynamics Research Corp., Wilmington, Massachusetts

CHARACTERISTICS OF THE BOUNDARY-LAYER EQUATIONS OF THE MINIMUM TIME-TO-CLIMB PROBLEM 807
Mark D. Ardema, Ames Research Center, NASA, Moffett Field, California

ON THE CONTROLLABILITY OF SINGULARLY PERTURBED NONLINEAR SYSTEMS 818
P. Sannuti, Rutgers University

SINGULARLY PERTURBED OPTIMAL CONTROL PROBLEMS 819
Paul Binding, University of Calgary, Alberta, Canada

THE TIME-OPTIMAL CONTROL OF A CLASS OF NONLINEAR SINGULARLY PERTURBED SYSTEMS 825
S. H. Javid, University of Illinois

STABILITY AND CONTROL OF STOCHASTIC SYSTEMS WITH WIDE-BAND NOISE DISTURBANCES 832
C. L. Blankenship, Case Western Reserve University and G. C. Papanicolaou, New York University

STOCHASTIC CONTROL OF LINEAR SINGULARLY PERTURBED SYSTEMS 841
A. H. Haddad and P. V. Kokotović, University of Illinois

ON LINEAR SINGULARLY PERTURBED SYSTEMS WITH STOCHASTIC INPUTS 851
Hassan K. Khalil, University of Illinois

LINEAR REGULAR DESIGN FOR STOCHASTIC SYSTEMS BY A MULTIPLE TIME SCALES METHOD 860
Demosthenis Teneketzis and Nils R. Sandell, Jr., Massachusetts Institute of Technology

IV-e: SOCIO-ECONOMIC SYSTEMS
Chairman: William B. Rouse, University of Illinois

A THEOREM CONCERNING PROPERTIES OF AN ARMS-RACE MODEL 875
I. W. Sandberg, Bell Laboratories, Murray Hill, New Jersey

A PIECEWISE-LINEAR INPUT-OUTPUT MODEL 881
Ming-Jeh Chien, Wayne State University

STATE SPACE MODELING OF A PLANT DISEASE EPIDEMIC . . 890
R. K. Varshney, Punjab Agricultural University, India

PROCESS DYNAMICS OF WHEAT GRAIN SECTOR IN PUNJAB STATE IN A 3-TIER HIERARCHY FOR WAREHOUSING AND TRANSPORTATION SYSTEM STUDIES 896
Ravindra Kumar and P. S. Satsangi, Indian Institute of Technology, New Delhi, India

ON INTEGRO-DIFFERENTIAL EQUATIONS ARISING IN POPULATION DYNAMICS AND CONTROL SYSTEMS 897
A. S. C. Sinha, Indiana Institute of Technology

SOCIO-ECONOMIC SYSTEMS MODELLING AND ANALYSIS OF A REAL-LIFE COMMUNITY HEALTH CARE SYSTEM 905
L. C. Agarwal and P. S. Satsangi, Indian Institute of Technology, New Delhi, India and J. B. Ellis, York University, Downsview, Canada

ENVIRONMENTAL - ECONOMIC SYSTEMS MODELLING 915
D. K. Bhargava and P. S. Satsangi, Institute of Technology, New Delhi, India and J. B. Ellis, York University, Downsview, Canada

V-a: ESTIMATION AND DETECTION
Chairman and Organizer: V. David VandeLinde, Johns Hopkins University

MAXIMUM-DISTANCE QUANTIZATION FOR DETECTION 925
H. V. Poor and J. B. Thomas, Princeton University

A STOCHASTIC FRAMEWORK FOR OPTIMAL CURVE FITTING . . 935
Howard L. Weinert, Johns Hopkins University and Gursharan S. Sidhu, State University of New York at Buffalo

ON APPROXIMATIONS TO CONDITIONAL MEAN ESTIMATES . . . 944
A. Fogel and S. C. Schwartz, Princeton University

SIGNAL DETECTION BASED ON QUANTIZED DATA 954
Saleem A. Kassam, University of Pennsylvania

MAXIMUM LIKELIHOOD SEQUENCE ESTIMATION FOR RANDOMLY DISPERSIVE CHANNELS 963
Robert E. Morley, Jr. and Donald L. Snyder, Washington University

V-b: SENSITIVITY AND IDENTIFICATION
Chairman: R. E. Klein, University of Illinois

COMPARATIVE SENSITIVITY AND ABSOLUTE INVARIANT COMPENSATORS 964
R. M. De Santis and Serge Lefebvre, Ecole Polytechnique, Montreal, Canada

CLOSED-LOOP STABILITY OF LINEAR QUADRATIC OPTIMAL SYSTEMS IN THE PRESENCE OF MODELING ERRORS 974
Mitsuhiko Toda, Rajnikant Patel and Banavar Sridhar, Ames Research Center, NASA, Moffett Field, California

ON THE ROBUST CONTROL OF A TIME DELAY SERVOMECHANISM PROBLEM 983
A. Goldenberg, Spar Aerospace Ltd, Toronto, Canada and E. J. Davison, University of Toronto

IDENTIFICATION OF MULTIVARIABLE SYSTEMS: A CRITICAL SURVEY 998
N. S. Rajbman, USSR Academy of Sciences, Moscow, U.S.S.R. and N. K. Sinha, McMaster University, Hamilton, Canada

IDENTIFICATION OF CERTAIN DYNAMIC SYSTEMS, FORCED BY DETERMINISTIC AND STOCHASTIC INPUTS, GIVEN THE INPUT AND NOISY OUTPUT MEASUREMENTS 1007
Earl D. Eyman and Ezekiel A. Adedotun, University of Iowa

OPTIMAL INPUT SIGNALS FOR IDENTIFICATION OF DISTRIBUTED PARAMETER SYSTEMS: A FREQUENCY-DOMAIN APPROACH 1016
B. Kuszta, Technical University of Warsaw, Warsaw, Poland and N. K. Sinha, McMaster University, Hamilton, Canada

AN EXTENSION OF MODIFIED FINAL PREDICTION ERROR CRITERION TO TESTING THE ORDERS OF MULTIVARIATE ARMA PROCESSES AND MULTIVARIABLE SYSTEMS 1023
Ezekiel A. Adedotun and Earl D. Eyman, University of Iowa

V-c: BROADBAND AMPLIFIERS
Chairman: Wataru Mayeda, University of Illinois

A NEW APPROACH TO BROAD BANDING 1033
Herbert J. Carlin, Cornell University

SMALL-SIGNAL THIRD-ORDER DISTORTION ANALYSIS OF TRANSISTOR AMPLIFIERS 1042
J. D. Cobb and T. N. Trick, University of Illinois

V-d: COMPUTATIONAL METHODS
Chairman: John E. Prussing, University of Illinois

GLOBAL MINIMIZATION OF POLYNOMIALS OF N VARIABLES WITH CONTROL APPLICATIONS 1052
W. E. Gesing and E. J. Davison, University of Toronto

ON THE BEHAVIOR OF OPTIMAL SAMPLED DATA REGULATORS DRIVEN BY STOCHASTIC INPUT 1063
Dinesh A. Keskar and Gary L. Slater, University of Cincinnati

OPTIMAL SAMPLING FOR DATA ACQUISITION SYSTEMS . . . 1064
Ronald L. Van Wieren, Lear Siegler, Inc., Grand Rapids, Michigan and Robert A. Schlueter, Michigan State University

SOLUTION OF MULTISENSOR TRACKING PROBLEMS ON ASSOCIATIVE PROCESSORS 1074
Charles L. Morefield, Orincon Corporation, La Jolla, California

LARGE SCALE LINEAR DYNAMIC PROGRAMMING USING THE METHOD OF MULTIPLIERS 1084
Lawrence E. Berman, The Analytic Sciences Corporation, Reading, Massachusetts

A COMPUTER PROGRAM TO FIND THE KERNEL OF A POLYNOMIAL OPERATOR 1091
R. R. Gejji, University of Notre Dame

ANALYSIS AND SYNTHESIS OF DYNAMIC SYSTEMS VIA BLOCK PULSE FUNCTIONS 1101
P. Sannuti, Rutgers University

V-e: DIGITAL SYSTEMS II
Chairman: E. S. Davidson, University of Illinois

A MULTI-MICROPROCESSOR ARCHITECTURE FOR SOR-TYPE ALGORITHMS 1102
Gerard G. L. Meyer, North Carolina State University

THE "GURU" - A MICROPROCESSOR SOLUTION 1113
Orin E. Marvel, Rao Aysola and Frank Zuhde, Honeywell, West Covina, California

TECHNIQUES FOR REDUCING PROCEDURAL DEPENDENCY IN A PROGRAM 1124
Lawrence S. Cheung, Marquette University and Frederick J. Mowle, Purdue University

MULTI-VALUED ASYNCHRONOUS SEQUENTIAL MODULES 1131
Anthony S. Wojcik, Illinois Institute of Technology

A GENERAL METHOD TO IMPLEMENT TERNARY CIRCUITS WITH BINARY COMPONENTS 1140
Daniel Etiemble, University of Paris, France and Michel Israel, CNAM Paris, France

AN INTRODUCTION TO DISCRETE TIME MARTINGALES AND FILTERING PROBLEMS

J. H. VAN SCHUPPEN
Department Systems Science & Mathematics
Washington University
Box 1040
St. Louis, Missouri 63130

Abstract

In this paper we present some concepts of discrete time martingale theory, including the recent developed concepts of semi-martingale decomposition, quadratic variation processes, martingale transforms and martingale representation.

We formulate the filtering problem and present the innovations method and the Bayesian method to solve this problem using martingale theory. Some examples are given.

1. INTRODUCTION

The purpose of this paper is to present an introduction to discrete time martingales and to show the application of this theory to filtering problems.

Martingales are a class of stochastic processes. They have been developed as a mathematical model for sequential gambling processes. The classical results of martingale theory center on the question of convergence of the process and technicalities related to this question. Recent developments in martingale theory are concerned with decompositions and transforms or stochastic integrals. It are these concepts that are of interest in applications.

Over the last several years publications have appeared that use martingale theory to derive solutions to stochastic problems in information and control theory. This approach has been very fruitfull, and we now have a general unified theory that solves many known stochastic problems. However there are examples of problems that can not be handled by martingale theory. We emphasize that this general theory starts from a general model and derives general results. To find explicit solutions one needs to specify the distributions of the particular processes.

What we will present in this paper is a simple introduction to discrete time martingales and the application of this theory to filtering problems. For precise statements and proofs we refer the reader to [2,3,10]. Related references are [4,8,9]. References on discrete time martingales are [1,5,6,7]. In the sections 2, 3 and 4 we discuss processes and martingales. In the sections 5 and 6 we present a discussion of the filtering problem.

Some of the research, reported on in this paper, is joint work with Dr. Brémaud, who is with the IRIA Laboratory, Paris, France.

2. DEFINITIONS AND NOTATIONS

We introduce some concepts from probability theory.

Let (Ω,F,P) denote a probability space. All variables and processes will take values in the set of the real numbers, together with the Borel σ-algebra.

The conditional expectation of an integrable random variable x given a sub-σ-algebra $G \subset F$, is a random variable $E[x|G]$ that is G measurable and such that for all $A \epsilon G$,

$$E[I_A \, x] = E[I_A \, E[x|G]]$$

Standard properties of conditional expectation can be found in [1,5,7].

Stochastic processes

Let $T \triangleq \{0,1,2,...t_1\}$ be the discrete time set. A stochastic process is denoted by $(x_t, t\epsilon T)$. The associated increment process $(\Delta x_t, t\epsilon T)$ is defined by $\Delta x_0 \triangleq x_0$, $\Delta x_t \triangleq x_t - x_{t-1}$ for t > 0. Then $x_t = \sum_{s=0}^{t} \Delta x_s$.

The Gaussian white noise process $(\Delta w_t, t\epsilon T)$ is defined to be an independent collection, and for all $t\epsilon T$ Δw_t has a Gaussian distribution with $E(\Delta w_t) = 0$. It will be called standard iff $\Delta w_0 = 0$ and $E[(\Delta w_t)^2] = 1$ for all $t\epsilon T$.

A counting process $(n_t, t\epsilon T)$ is a process that satisfies $n_0 = 0$ and for $t > 0$, $\Delta n_t = 0$ or 1. A Bernoulli process with parameter $p\epsilon R$, $0 \leq p \leq 1$, is a counting process $(n_t, t\epsilon T)$, with $(\Delta n_t, t\epsilon T)$ an independent collection and for all $t\epsilon T$, Δn_t has a Bernoulli distribution: $P(\Delta n_t = 1) = p$, $P(\Delta n_t = 0) = 1 - p$.

If $(x_t, t\epsilon T)$ is a process, then $F^x_t \triangleq F(x_s, \forall s \leq t)$ denotes the σ-algebra generated by the process x up to $t\epsilon T$. Then $(F^x_t, t\epsilon T)$ is a σ-algebra family that is increasing with $t\epsilon T$. A process $(x_t, F_t, t\epsilon T)$ is adapted to a σ-algebra family iff for all $t\epsilon T$ x_t is F_t measurable. A process $(x_t, F_t, t\epsilon T)$ is predictable iff for all $t\epsilon T$, x_{t+1} is F_t measurable, and $x_0 \equiv E(x_0)$. Interpretation: F_t predicts x_{t+1}.

3. MARTINGALES

In this section we define a martingale and some related concepts.

Definition 3.1. A process $(m_t, F_t, t\epsilon T)$ is a martingale iff:

(i) $(m_t, F_t, t\epsilon T)$ is adapted, (ii) $E|m_t| < \infty$ for all $t\epsilon T$,

(iii) $E(\Delta m_{t+1} | F_t) = 0$ a.s. for all $t\epsilon T$.

Notation: $(m_t, F_t, t\epsilon T) \; \epsilon \; M$

A martingale is a mathematical model for a gambling process. Condition (iii) expresses that the average future gain given current information, is zero. We give two examples of martingales:

- If $(\Delta w_t, t\epsilon T)$ is standard Gaussian white noise, then $(w_t, F^w_t, t\epsilon T) \; \epsilon \; M$. Proof: $E(\Delta w_{t+1} | F^w_t) = E(\Delta w_{t+1}) = 0$ by the independence of $(\Delta w_t, t\epsilon T)$.
- If $(n_t, t\epsilon T)$ is a Bernoulli process with parameter p, then $(n_t - pt, F^n_t, t\epsilon T) \; \epsilon \; M$. Proof: $E(\Delta n_{t+1} - p | F^n_t) = E(\Delta n_{t+1} - p) = 0$.

Definition 3.2. A process $(x_t, F_t, t\epsilon T)$ will be called a semi-martingale iff it has a decomposition $x = a + m$ where $(a_t, F_t, t\epsilon T)$ is an integrable predictable process and $(m_t, F_t, t\epsilon T) \; \epsilon \; M$ with $E(m_0) = 0$.

Proposition 3.3. The process $(x_t, F_t, t \in T)$ has an unique semi-martingale decomposition iff it is an adapted integrable process.

Hint of proof: $\Delta x_t = E[\Delta x_t | F_{t-1}] + (\Delta x_t - E[\Delta x_t | F_{t-1}]) = \Delta a_t + \Delta m_t$.

For details see [2, th. 2.1]. This equation clearly shows the essential point of the semi-martingale decomposition; it projects Δx_t on F_{t-1} to form the predictable part $\Delta a_t = E[\Delta x_t | F_{t-1}]$. The difference is Δm_t. Applications of the semi-martingale decomposition:

- Let $(x_t, t \in T)$ be a Gauss-Markov process, $E(x_t) = 0$ for all $t \in T$. Then it has the representation

  $$x_{t+1} = \alpha_t x_t + \Delta w_{t+1}, \qquad x_0 = \Delta w_0$$

  where $(\Delta w_t, t \in T)$ is Gaussian white noise, non-standard and for all $t \in T$, $\alpha_t \in R$.

- Let $(n_t, F_t, t \in T)$ be an adapted counting process. Define the rate process $\lambda_t = E[\Delta n_{t+1} | F_t]$ for all $t \in T$, $\lambda_0 = 0$. Then we have the semi-martingale decomposition

  $$\Delta n_{t+1} = \lambda_t + \Delta m_{t+1}.$$

Definition 3.4. Let $(x_t, F_t, t \in T)$ and $(y_t, F_t, t \in T)$ be processes. Then we define the processes

$$[x,]_t \triangleq \sum_{s=0}^{t} (\Delta x_s)(\Delta y_s)$$

$$<x,y>_t \triangleq E(x_0 y_0) + \sum_{s=1}^{t} E[\Delta x_s \Delta y_s | F_{s-1}]$$

called respectively the optional and predictable quadratic covariation process. The $<x,y>$ process exists if $E|\Delta x_t \Delta y_t| < \infty$ for all $t \in T$.

Examples: If $(\Delta w_t, t \in T)$ is standard Gaussian white noise, then $(w_t, F_t^w, t \in T) \in M$ with $<w,w>_t = t$.

- If $(n_t, t \in T)$ is a Bernoulli process with parameter p, then $(n_t - pt, F_t^n, t \in T) \in M$ with $<n_t - pt, n_t - pt>_t = p(1-p)t$.

4. TRANSFORMS

A transform is the discrete time equivalent of a stochastic integral. We state the definition of a transform and some of its properties.

Definition 4.1. The transform of a process $(x_t, t \in T)$ by a process $(h_t, t \in T)$ is the process $((h.x)_t, t \in T)$ defined by

$$(h.x)_t = \sum_{s=0}^{t} h_s \Delta x_s \quad \text{for all } t \in T.$$

Proposition 4.2. Properties of the transform operation.

a. Linearity $((h_1 + h_2).x) = (h_1.x) + (h_2.x)$

$$((\alpha h).x) = \alpha(h.x) \qquad \forall \alpha \in R$$

b. $[(h.x),y] = (h.[x,y])$

c. $<(h.x),y> = (h.<x,y>)$

if the processes $<x,y>$ and $<(h.x),y>$ exist, and if h is predictable.

Proposition 4.3. The martingale transform

If $(m_t, F_t, t\varepsilon T) \varepsilon M$, $(h_t, F_t, t\varepsilon T)$ is a predictable process and $E|h_t \Delta m_t| < \infty$ for all $t\varepsilon T$, then $((h.m)_t, F_t, t\varepsilon T) \varepsilon M$.

Proof. $E[\Delta(h.m)_t | F_{t-1}] = h_t \, E[\Delta m_t | F_{t-1}] = 0$.

Proposition 4.4. The product rule.

For processes $(x_t, t\varepsilon T)$ and $(y_t, t\varepsilon T)$ we have

$$x_t y_t = \sum_{s=1}^{t} x_{s-1} \Delta y_s + \sum_{s=1}^{t} y_{s-1} \Delta x_s + [x,y]_t$$

A type of exponential formula process can be constructed also, which is of interest in the study of detection problems. (see [9]).

The following martingale representation theorems are extremely important and powerful results, and the key to the application of martingale theory.

Theorem 4.5. Let $(w_t, F_t^w, t\varepsilon T) \varepsilon M$ with $E[(\Delta w_t)^2] > 0$ for all $t\varepsilon T$, $w_0 = 0$, and $(m_t, F_t^w, t\varepsilon T) \varepsilon M$. Assume that $(m_t, w_t, t\varepsilon T)$ is a collection of jointly Gaussian random variables. Then m has a representation as a transform

$$m_t = E(m_0) + \sum_{s=1}^{t} h_s \, \Delta w_s$$

where $h_t = E(m_t \, \Delta w_t)/E[(\Delta w_t)^2]$ for all $t\varepsilon T$.

Proof. [2, th. 3.1].

Theorem 4.6. Let $(n_t, t\varepsilon T)$ be a counting process.

Let $(\lambda_t, F_t^n, t\varepsilon T)$ be its rate process, $\lambda_t = E[\Delta n_{t+1} | F_t^n]$, $\lambda_0 = 0$.

Then $(\sum_{s=0}^{t} (\Delta n_s - \lambda_{s-1}), F_t^n, t\varepsilon T) \varepsilon M$. Furthermore any $(m_t, F_t^n, t\varepsilon T)$ has a representation as a transform

$$m_t = E(m_0) + \sum_{s=1}^{t} h_s (\Delta n_s - \lambda_{s-1})$$

for a predictable process $(h_t, F_t^n, t\varepsilon T)$ satisfying $E[|h_t| \lambda_{t-1}(1-\lambda_{t-1})] < \infty$ for all $t\varepsilon T$.

Proof. [2, Cor. 3.5]

A more general result has been obtained [2, th. 3.4].

5. THE FILTERING PROBLEM

We formulate the filtering problem as follows: Given the process $(y_t, t\varepsilon T)$ that is observed. Given a process $(x_t, t\varepsilon T)$ that is not observed. The relation between x and y is as of yet unspecified. The filtering problem is to estimate for all $t\varepsilon T$ the variable x_t given all observations up to time $t\varepsilon T$.

Suppose that we further specify that the estimate must be according to the minimum variance criterion. Then the optional estimate is the process $(\hat{x}_t, F_t^y, t\varepsilon T)$, $\hat{x}_t \overset{\Delta}{=} E[x_t | F_t^y]$. In principle this solves the filtering problem.

We now define the filter realization problem as to obtain a system realization or filter that produces the $\hat{x}$ process given the y process. The only interesting filters are those that are finite dimensional or recursive, since such filters can be implemented.

We discuss two basic methods for the filter realization problem, namely the innovations method and the Bayesian method.

The innovations method

We define the innovations process and obtain a filter representation according to the innovations method.

Definition 5.1. Give the observed process $(y_t, t\varepsilon T)$ and assume that it is integrable. Define the innovations process $(\bar{w}_t, t\varepsilon T)$ by $\bar{w}_0 \overset{\Delta}{=} y_0 - E(y_0)$, $\Delta\bar{w}_t \overset{\Delta}{=} y_t - E[y_t | F_{t-1}^y]$ for $t > 0$.
Remark that this is essentially the semi-martingale decomposition 3.4.

Proposition 5.2. The innovations process $(\bar{w}_t, t\varepsilon T)$ satisfies:

a. $(\bar{w}_t, F_t^y, t\varepsilon T) \ \varepsilon\ M$,

b. $F_t^{\bar{w}} = F_t^y$ for all $t\varepsilon T$.

Definition 5.3. Assume that the integrable processes x and y have the following semi-martingale decomposition with respect to an increasing family $(F_t, t\varepsilon T)$

$$x_{t+1} = f_t + \Delta v_{t+1}, \quad x_0 = v_0$$

$$y_{t+1} = h_t + \Delta w_{t+1}$$

where $(v_t, F_t, t\varepsilon T) \ \varepsilon\ M$, $(w_t, F_t, t\varepsilon T) \ \varepsilon\ M$, $w_0 = 0$, and $(f_t, F_t, t\varepsilon T)$ and $(h_t, F_t, t\varepsilon T)$ are integrable adapted processes.
Remark. If the model for the processes x and y is explicitly given, then the identification with the above model can easily be found.

Proposition 5.4. With respect to the family $(F_t^y, t \varepsilon T)$, generated by the observed process y, we have the representation

$$\hat{x}_{t+1} = \hat{f}_t + \Delta \bar{v}_{t+1}, \quad \hat{x}_0 = E(x_0)$$

$$y_{t+1} = \hat{h}_t + \Delta \bar{w}_{t+1}$$

where $\hat{x}_t = E[x_t | F_t^y]$, similarly $\hat{f}_t$ and $\hat{h}_t$, and $(\bar{v}_t, F_t^y, t \varepsilon T) \varepsilon M$, $(\bar{w}_t, F_t^y, t \varepsilon T) \varepsilon M$. $\bar{w}$ is the innovations process.

For the well known linear model with Gaussian white noise disturbances, the discrete time Kalman filter can now be obtained. This follows, with some calculations, from proposition 5.4 and theorem 4.5. The details are given in [9]. We note that the Kalman filter is a finite dimensional filter realization.

We now state a result for counting process observations.

Definition 5.5. Consider the special case of definition 5.3.

$$\lambda_{t+1} = f_t + \Delta v_{t+1}, \quad \lambda_0$$

$$\Delta n_{t+1} = \lambda_t + \Delta w_{t+1}$$

where the observed process is the counting process $(n_t, t \varepsilon T)$, and where $\lambda_t = E[\Delta n_{t+1} | F_t]$, $(\lambda_t, F_t, t \varepsilon T)$ is the associated rate process with the representation given. Thus $(v_t, F_t, t \varepsilon T) \varepsilon M$ and $(w_t, F_t, t \varepsilon T) \varepsilon M$.

Theorem 5.6. For the processes of definition 5.5 we have

$$\hat{\lambda}_{t+1} = \hat{f}_t + (\hat{q}_t^{f,\lambda} + \hat{q}_t^{v,w}) \hat{\lambda}_t^{-1} (1 - \hat{\lambda}_t)^{-1} (\Delta n_{t+1} - \hat{\lambda}_t),$$

$$\hat{\lambda}_0 = E(\lambda_0),$$

where $$\hat{q}_t^{f,\lambda} = E[(f_t - \hat{f}_t)(\lambda_t - \hat{\lambda}_t) | F_t^n]$$

$$\hat{q}_t^{v,w} = E[\Delta <v,w>_{t+1} | F_t^n]$$

Proof outline: By proposition 5.4.

$$\hat{\lambda}_{t+1} = \hat{f}_t + \Delta \bar{v}_{t+1}, \quad \hat{\lambda}_0 = E(\lambda_0)$$

$$\Delta n_{t+1} = \hat{\lambda}_t + \Delta \bar{w}_{t+1}$$

It is then easy to show that

$$\Delta\langle\bar{w},\bar{w}\rangle_t = \hat{\lambda}_{t-1}(1-\hat{\lambda}_{t-1})$$

$$\Delta\langle\bar{v},\bar{w}\rangle_t = \hat{q}^{f,\lambda}_{t-1} + \hat{q}^{v,w}_{t-1}$$

Since $(\bar{v}_t, F^n_t, t\varepsilon T) \ \varepsilon \ M$, by theorem 4.6.

$$\Delta\bar{v}_t = k_t \ \Delta\bar{w}_t = k_t \ (\Delta n_t - \hat{\lambda}_{t-1})$$

$$\Delta\langle\bar{v},\bar{w}\rangle_t = k_t \ \Delta\langle\bar{w},\bar{w}\rangle_t$$

The expression for k_t can then be determined. The details can be found in [3, prop. 3.6; 9].
Remark. Theorem 5.6 was first published by Segall [8], and was also obtained in [3].

We emphasize that 5.6 is not a filter realization since $\hat{f}$, $\hat{q}^{f,\lambda}$ and $\hat{q}^{v,w}$ are not specified in terms of a filter state. However a finite dimensional filter realization can be obtained if it is assumed that the rate process λ is a finite state Markov process (see [2]).

We now state the general result.

Theorem 5.7. Given the processes $(x_t, t\varepsilon T)$ and $(y_t, t\varepsilon T)$ as defined in 5.3. Then the minimum variance estimate $(\hat{x}_t, t\varepsilon T)$ satisfies the filter representation

$$\hat{x}_{t+1} = \hat{f}_t + \int_R k(t+1,u) \ \tilde{y}(t+1,du)$$

where: 1. $(\tilde{y}(t,A), F^y_t, t\varepsilon T, A\varepsilon B)$ is the innovations kernel of $(y_t, t\varepsilon T)$,

2. for any bounded real valued predictable process $(g(t,u), F^y_t, t\varepsilon T, u\varepsilon R)$ we have

$$E[(f_{t-1} - \hat{f}_{t-1} + \Delta v_t)(\int_R g(t,u) \ \tilde{y}_m(t,du))]$$

$$= E[\int_R k(t,u)[g(t,u) - \int_R g(t,z) \ \tilde{y}_{tr}(t,dz)] \ \tilde{y}_{tr}(t,du)]$$

Proof. [3, th. 3.4]

Theorem 5.7 gives the general filter representation. We emphasize again that 5.7 is a representation only, and not a filter realization. This is because $\hat{f}$ and k are not yet specified in terms of a state of the filter. The Kalman filter and the filter representation 5.6 for counting process observations can be deduced from theorem 5.7.

The main disadvantage of the innovations method is that it does not help us to determine the conditional distribution of the unobserved process. The Bayesian method seems more promising in this respect.

6. THE BAYESIAN METHOD

We shortly summarize the Bayesian method to solve filtering problems. One performs a measure transformation, if that is possible, such that under the new measure the observed process and the process to be estimated are independent. The classical estimation problem in statistics can then be solved directly. To solve the filtering problem we have developed a procedure which yields the well known Kalman filter and theorem 6.2 below. For details see [10].

Definition 6.1. Consider the processes specified by

1. $(n_t, t \varepsilon T)$ takes values in $Z_+ \cup \{0\}$, $n_0 = 0$, and

$$E[I_{(\Delta n_{t+1} = k)} | F^{\lambda}_{t_1} \vee F^{n}_{t}] = \frac{\lambda_t^k}{k!} e^{-\lambda_t} \quad \forall k \varepsilon Z_+ \cup \{0\},\ \forall t \varepsilon T,$$

2. $\lambda_{t+1} = \alpha\lambda_t + \Delta v_{t+1}$, $\lambda_0 = \Delta v_0$,

3. $(\Delta v_t, t \varepsilon T)$ is a collection of independent random variables and for all $t \varepsilon T$, Δv_t has a Gamma distribution with parameters $r_t, b_t \varepsilon R_+$, $r_t > 0$, $b_t > 0$.

4. $b_{t+1} = \frac{1}{\alpha}(1 + b_t)$, $\alpha > 0$.

These processes have the representation

$$\lambda_{t+1} = \alpha\lambda_t + \Delta v_{t+1}, \quad \lambda_0 = \Delta v_0$$

$$\Delta n_{t+1} = \lambda_t + \Delta w_{t+1}$$

with $(w_t, F^{\lambda}_{t_1} \vee F^{n}_{t}, t \varepsilon T) \varepsilon M$.

Theorem 6.2. For the model 6.1 the optimal estimate has the conditional characteristic function

$$E[e^{iu\lambda_t} | F^n_t] = \left(1 - \frac{iu}{b_t}\right)^{-\sum_{s=0}^{t} (\Delta n_s + r_s)}$$

which is of Gamma type, where $(n_t, t \varepsilon T)$ is a sufficient statistic.

A filter realization in terms of $(\hat{\lambda}_t, t \varepsilon T)$ can be given by

$$\hat{\lambda}_{t+1} = \alpha\hat{\lambda}_t + \frac{r_{t+1}}{b_{t+1}} + b^{-1}_{t+1}(\Delta n_{t+1} - \hat{\lambda}_t)$$

$$\hat{\lambda}_0 = r_0/b_0$$

$$b_{t+1} = \frac{1}{\alpha}(b_t+1),\ b_0$$

Proof. [10, th. 4.5].

REFERENCES

1. Ash, R.B., "Real analysis and probability", Academic Press, N.Y., 1972.
2. Brémaud, P.M., Van Schuppen, J.H., "Discrete time processes: Preliminaries and Part I: Martingale Calculus and Innovations Kernels", preprint, June 1976.
3. Brémaud, P.M., Van Schuppen, J.H., "Discrete time processes: Estimation", preprint.
4. Brémaud, P.M., "Countable state space signals and multivariate point process observations: Application to ALOHA", preprint, June 1976.
5. Meyer, P. A., "Probability and Potentials", Blaisdell, Waltham, Mass., 1966.
6. Meyer, P.A., "Martingales and Stochastic Integrals I", Lecture Notes in Math., vol. 284, Springer, Berlin, 1972.
7. Neveu, J., "Martingales à Temps Discrete", Masson, France, 1972; English translation: "Discrete parameter martingales", North-Holland Publ. Comp., Amsterdam, 1975.
8. Segall, A., "Recursive estimation from discrete-time point proceses", IEEE-T-INFO, 22, 4, (July, 1976), pp. 422-431.
9. Van Schuppen, J.H., "Notes on Estimation and Stochastic Control", St. Louis, Mo., Spring 1976.
10. Van Schuppen, J. H., "A study of estimation and filtering problems by the Bayesian method", preprint, June 1976.

OPTIMAL, CAUSAL, SIMULTANEOUS DETECTION AND ESTIMATION OF RANDOM SIGNAL FIELDS IN GAUSSIAN NOISE FIELD

T. T. KADOTA
Bell Telephone Laboratories, Murray Hill, New Jersey

ABSTRACT

We present two methods of simultaneously detecting a random signal field and estimating its Markovian parameters in the presence of white Gaussian noise field. The first method provides the Neyman-Pearson decision rule for determining the absence-or-presence of the signal and the maximum a posteriori likelihood estimators of the parameters. The second provides the Bayes decision rule for the absence-or-presence of the signal and the minimum mean-square-error estimators of the parameters. Both methods allow continual, causal detection and estimation based on the continuous space-time data, and their detection and estimation statistics are given in terms of a single statistic, which is given as the solution of a certain stochastic integral equation. Furthermore, we develop an approximate scheme of recursively generating this statistic by using spatially and temporally sampled data.

INTRODUCTION

We present some highlights of the results reported in another paper [0]. Consider the problem of simultaneously detecting a random signal field and estimating its Markovian parameters in the presence of white Gaussian noise field. We require that both detection and estimation be made on the observed data which are spatially and temporally continuous. Furthermore, they are to be performed continually in time, and, at each instant of time, the operation be casual and the result be optimum according to some criteria.

Separate detection and estimation of Markovian signals in Gaussian noise have been studied by many, and recursive procedures have been obtained in various forms [1-9]. Joint detection and estimation, on the other hand, has not been treated as extensively. Starting with Middleton and Esposite [10,11], there have been several authors [12-18] who dealt with the joint problem. They all formulated it as (Bayes) estimation problem though they differed in the cost-functions and the observation data they used. The most general model of the observation data was a continuous-time vector process, and the mathematical derivation was rather formal in this case.

We give a decision-oriented formulation as well as the estimation-oriented. In the former, the Neyman-Pearson criterion is used for detection performance and the maximum a posteriori likelihood method is used for estimation. In the latter, the minimum average error-probability criterion for detection and the minimum mean-square-error for estimation are used. Both formulations are based on continuous space-time data, that is, the signals and noise considered are continuous-parameter random fields. A causal, recursive solution is given for each formulation, and approximate schemes are provided for practical implementation. In addition to establishing a general mathematical theory, we treat a concrete problem of detecting a narrow-band signal in white Gaussian noise and estimating the direction of the signal arrival and the signal carrier frequency. For the signal model, we assume that the direction is a random variable and the carrier frequency is a Markov process. When the detection and estimation are performed continually in time, certain "tracks" of the signal-arrival direction and of the carrier frequency are obtained as time goes on. When the signal is not detected, a gap appears in each of these tracks. The direction track can be used for locating the signal source and the frequency track for identifying the type of the signal source. Thus, this problem can be viewed as a simple example of signal detection, localization and identification.

MAIN RESULTS AND DISCUSSIONS

The signal model we adopt is the following:

$$\tilde{v}(t)\cos \phi(\underline{r},t) + \hat{v}(t)\sin \phi(\underline{r},t), \quad \underline{r} \in A \subset R^3, \quad 0 \leq t < \infty,$$

where

$$\phi(\underline{r},t) = 2\pi f_0(t+\underline{k}\cdot\underline{r}/c) + 2\pi \int_0^{t+\underline{k}\cdot\underline{r}/c} f(s)ds + \phi_0,$$

$\tilde{v}(\cdot)$, $\hat{v}(\cdot)$ are random amplitudes of the in-phase and the quadrature components of the plane-wave signal, which are independent, identically distributed Gauss-Markov processes with mean 0,

f_0 is the average frequency of the signal,
$\underline{k}$ is the random directional vector of the plane-wave,
$\underline{r}$ is the coordinate vector of an observation point in A,
A is a 3-dimensional observation region in R^3, $R = (-\infty,\infty)$,
c is the speed of propagation of the signal,
$f(\cdot)$ is the random frequency of the signal, which is a stationary Markov process with mean 0,
ϕ_0 is the random phase, which is uniformly distributed on $[0,2\pi)$.

The noise model we adopt is a 4-dimensional white Gaussian field. For a precise mathematical treatment, we interpret it as a symbolic derivative of the Wiener process with a 4-dimensional parameter $(\underline{r},t)$. We use the random set-function version of the Wiener process $\{W(B),\ B \subset R^3 \times [0,\infty)\}$ where, for any sequence of Borel sets $B_1,\ldots,B_n$ in $R^3 \times [0,\infty)$, $W(B_1),\ldots,W(B_n)$ are jointly Gaussian distributed with mean $E\,W(B_j) = 0$ and covariance $E\,W(B_i)W(B_j) = \mu(B_iB_j)$, $i,j = 1,\ldots,n$, and μ is the Lebesgue measure in 4-dimension. We assume that the signal parameters and the noise are mutually independent.

The problem of simultaneous detection and estimation is to find a rule for deciding whether or not the signal is present, and to obtain estimators of the signal parameters $\underline{k}$ and $f(t)$, based on the same observed data. We formulate this problem in two different ways. Put

$$\theta_i(t) = k_i, \quad i = 1,2 \quad \text{where} \quad \underline{k} = (k_1,k_2,k_3), \quad k_1^2+k_2^2+k_3^2 = 1,$$

$$\theta_3(t) = 2\pi f(t), \quad \theta_4(t) = \left(2\pi \int_0^{t+\underline{k}\cdot\underline{r}/c} f(s)ds + \phi_0\right)_{2\pi},$$

$$\theta_5(t) = \tilde{v}(t), \quad \theta_6(t) = \hat{v}(t),$$

and

$$\underline{\theta}(t) = (\theta_1(t),\ldots,\theta_6(t)), \quad \Theta = [0,1]\times[0,1]\times[0,\infty)\times[0,2\pi)\times R\times R,$$

$$z(\underline{r},t,\underline{\theta}(t)) = \tilde{v}(t)\cos\phi(\underline{r},t) + \hat{v}(t)\sin\phi(\underline{r},t),$$

where the symbol $(\cdot)_{2\pi}$ reads "modulo 2π".

Formulation I (decision-oriented)

Let P_0 and P_1 be two probability measures induced, in the space $\mathcal{Y}$ of all real-valued set-functions $\{Y(B),\ B \subset A \times [0,\infty)\}$, by $\{W(B),\ B \subset A \times [0,\infty)\}$ and $\{\int_B z(\underline{r},t,\underline{\theta}(t))d\underline{r}dt+W(B),\ B \subset$

$A \times [0,\infty)\}$ respectively. The observable in this formulation is $Y \in \mathcal{Y}$, and a decision rule is a partition $(\mathcal{Y}_0, \mathcal{Y}_1)$ of $\mathcal{Y}$ with the interpretation that $Y \in \mathcal{Y}_0$ implies choice of P_0 and vise versa. Here P_0 and P_1 correspond to the signal-absent and the signal-present hypotheses. Then, there exists a likelihood ratio $\bar{\ell}_{A,t}$, which is a functional of $\{Y(B),\ B \subset A \times [0,t]\}$, between the "finite-observation-time" restrictions of P_0 and P_1. Thus, according to the Neyman-Pearson theorem, the following partition maximizes the detection probability for a given false-alarm probability ε:

$$\mathcal{Y}_0 = \{Y:\ \bar{\ell}_{A,t}(Y) < b_{A,t}\}, \qquad \mathcal{Y}_1 = \{Y:\ \bar{\ell}_{A,t}(Y) \geq b_{A,t}\}$$

where the threshold $b_{A,t}$ is determined by $P_0(\{Y:\ \bar{\ell}_{A,t}(Y) \geq b_{A,t}\}) = \varepsilon$.

Let $\underline{V}$ be the space of all vector-valued functions $\{\underline{v}(t),\ 0 \leq t < \infty\}$ where $\underline{v}(t) \in \boldsymbol{\Theta}$ for each t, and $P_{\underline{\Theta}}$ be the probability measure induced in $\underline{V}$ by $\{\underline{\Theta}(t),\ 0 \leq t < \infty\}$. Denote by $P_0 \times P_{\underline{\Theta}}$ and $P_1 \times P_{\underline{\Theta}}$ two product measures in the product space $\mathcal{Y} \times \underline{V}$ of pairs $(Y, \underline{v})$. Then, for each t, there exists a likelihood ratio $\bar{\ell}^{(1,2,3)}_{A,t}$ between the restrictions of $P_0 \times P_{\underline{\Theta}}$ and $P_1 \times P_{\underline{\Theta}}$, which is a functional of $\{Y(B),\ B \subset A \times [0,t]\}$ and also a function of $v_1(t)$, $v_2(t)$, and $v_3(t)$. We assume that k_1, k_2 and $f(t)$ have a joint probability density, which we denote by $p_{\Theta_i(t),i=1,2,3}(\cdot,\cdot,\cdot)$. Then, the maximum a posteriori likelihood estimators $\hat{k}_{1,A,t}(\cdot)$, $\hat{k}_{2,A,t}(\cdot)$, $\hat{f}_{A,t}(\cdot)$ of k_1, k_2 and $f(t)$ respectively are given by

$$\max_{\substack{0 \leq \eta_1, \eta_2 \leq 1 \\ -\infty < \eta_3 < \infty}} p_{\Theta_i(t),i=1,2,3}(\eta_1,\eta_2,\eta_3)\,\bar{\ell}^{(1,2,3)}_{A,t}(Y,(\eta_1,\eta_2,\eta_3))$$

$$= p_{\Theta_i(t),i=1,2,3}(\hat{k}_1(Y),\hat{k}_2(Y),\hat{f}(Y))\,\bar{\ell}^{(1,2,3)}_{A,t}(Y,(\hat{k}_1(Y),\hat{k}_2(Y),\hat{f}(Y))),$$

where the subscripts "A,t" of k_1, k_2 and f are temporarily suppressed for convenience.

Formulation II (estimation-oriented)

Introduce a random variable u which takes the value 0 or 1 and is independent of all random quantities $\tilde{v}(\cdot)$, $\hat{v}(\cdot)$, $\underline{k}$, $f(\cdot)$, $\emptyset_0$ and $W(\cdot)$. Define a new process $\{\tilde{X}(B),\ B \subset A \times [0,\infty)\}$ by

$$\tilde{X}(B) = \int_B uz(\underline{r},t,\underline{\Theta}(t))\, d\underline{r}dt + W(B).$$

The observable in this formulation is the sample function of $\{\tilde{X}(B),\ B \subset A \times [0,t]\}$. Then, a decision rule is a partition $(\mathcal{X}_0,\mathcal{X}_1)$ of the sample space of $\{\tilde{X}(B),\ B \subset A \times [0,t]\}$, and the average error-probability P_e is defined as

$$P_e(\mathcal{X}_0,\mathcal{X}_1) = P(u = 0,\ \tilde{X} \in \mathcal{X}_1) + P(u = 1,\ \tilde{X} \in \mathcal{X}_0).$$

It is shown that the following partition minimizes P_e:

$$\mathcal{X}_0 = \{\tilde{X}\colon\ E\{u|X_{A,t}\} < 1/2\}, \qquad \mathcal{X}_1 = \{\tilde{X}\colon\ E\{u|X_{A,t}\} \geq 1/2\}.$$

where $\tilde{X}_{A,t} = \{\tilde{X}(B),\ B \subset A \times [0,t]\}$. The minimum mean-square-error estimators of k_1, k_2 and $f(t)$ are, of course, given by

$$E\{k_i|\tilde{X}_{A,t}\}, \quad i = 1,2, \quad E\{f(t)|X_{A,t}\}.$$

It is seen that in Formulation I the optimum decision rule and estimators are specified in terms of the two likelihood ratios $\bar{\ell}_{A,t}(\cdot)$ and $\bar{\ell}^{(1,2,3)}_{A,t}(\cdot,\cdot,\cdot)$, $t \in [0,\infty)$; while in Formulation II they are specified in terms of the conditional expectations of u, k_1, k_2 and $f(t)$ given $\tilde{X}_{A,t}$ for each $t \in [0,\infty)$. Thus, the task of obtaining the optimum decision rules and the estimators is reduced to that of finding concrete expressions for the two likelihood ratios in Formulation I and the four conditional expectations in Formulation II.

Put

$$\Theta_0(t) = u, \quad \Theta_7(t) = \sup_{0 \leq s \leq t} [\tilde{v}^2(s) + \hat{v}^2(s)]^{1/2}$$

and add these components to the parameter vector $\underline{\Theta}(t)$, $0 \leq t < \infty$. Redenote the resultant vector by $\tilde{\underline{\Theta}}(t)$, and the corresponding state space by $\tilde{\underline{\boldsymbol{\Theta}}}$. Put

$$\tilde{z}(\underline{r},t,\tilde{\underline{\Theta}}(t)) = \Theta_0(t)\ z(\underline{r},t,\underline{\Theta}(t)).$$

Then it is shown, under certain minor conditions on z, that a stochastic integral equation

$$\tilde{\psi}(A \times [0,t],Y,\tilde{\underline{\eta}}) = 1 + \int_{A \times [0,t]} \int_{\tilde{\underline{\boldsymbol{\Theta}}}} \tilde{z}(\underline{r},s,\tilde{\underline{\xi}})\tilde{\psi}(A \times [0,s],Y,\tilde{\underline{\xi}}) \cdot P(\tilde{\underline{\Theta}}(s) \in d\tilde{\underline{\xi}} | \underline{\Theta}(t) = \tilde{\underline{\eta}})Y(d\underline{r} \times ds) \tag{1}$$

has a unique solution $\tilde{\psi}$ for almost every realization $(Y,\tilde{\underline{\eta}})$ of $W,\tilde{\underline{\Theta}}(t))$; furthermore,

$$\bar{\ell}_{A,t}(Y) = \int_{\tilde{\underline{\boldsymbol{\Theta}}}} \tilde{\psi}(A \times [0,t],Y,\tilde{\underline{\xi}})\Big|_{\xi_0=1} P(\tilde{\underline{\Theta}}(t) \in d\tilde{\underline{\xi}}), \tag{2}$$

$$\bar{\ell}^{(1,2,3)}_{A,t}(Y,(\eta_1,\eta_2,\eta_3)) = \int_{\tilde{\underline{\boldsymbol{\Theta}}}} \tilde{\psi}(A \times [0,t],Y,\tilde{\underline{\xi}})\Big|_{\xi_0=1,\xi_i=\eta_i,i=1,2,3} \cdot P((\Theta_4(t),\dots,\Theta_7(t)) \in d(\xi_4,\dots,\xi_7) | \Theta_3(t) = \eta_3) \tag{3}$$

and if we put

$$\hat{\theta}_{i,A,t}(\tilde{X}) = \Big[\int_{\tilde{\underline{\boldsymbol{\Theta}}}} \tilde{\psi}(A \times [0,t],\tilde{X},\tilde{\underline{\xi}})P(\tilde{\underline{\Theta}}(t) \in d\tilde{\underline{\xi}})\Big]^{-1} \cdot \int_{\tilde{\underline{\boldsymbol{\Theta}}}} \xi_i\tilde{\psi}(A \times [0,t],\tilde{X},\tilde{\underline{\xi}})P(\tilde{\underline{\Theta}}(t) \in d\tilde{\underline{\xi}}),\ i = 0,1,2,3, \tag{4}$$

then

$$E\{u|\tilde{X}_{A,t}\} = \hat{\theta}_{0,A,t}(\tilde{X}), \quad E\{k_i|\tilde{X}_{A,t}\} = \hat{\theta}_{i,A,t}(\tilde{X}), \quad i = 1,2,$$
$$E\{f(t)|\tilde{X}_{A,t}\} = \hat{\theta}_{3,A,t}(\tilde{X}). \tag{5}$$

Observe that all the statistics for detection and estimation in both formulations are given in terms of $\tilde{\psi}$. While the probability distribution $P(\tilde{\underline{\Theta}}(t) \in d\tilde{\underline{\xi}})$ of the parameters is required for determining these statistics, $\tilde{\psi}$ is the only quantity that depends on the observed data Y (or $\tilde{X}$). Thus, $\tilde{\psi}$ is the only statistic that must be continually up-dated as new data are observed, and this up-dating is done recursively as the integral equation indicates.

Introduction of Θ_7 is necessitated in order to assure uniqueness of the solution of the equation (1). Without uniqueness, specification of $\tilde{\psi}$ as the solution of (1) is incomplete (in fact, not specification at all). As a result, the vector process $\{\tilde{\underline{\Theta}}(t),\ 0 \leq t < \infty\}$ is no longer stationary since Θ_7 is not. Aside from the complication due to Θ_7, the equation (1) is in general difficult to solve, whether explicitly or in terms of feedback-network implementation. For practical application, therefore, it is desirable to have an approximate recursive formula for calculating $\tilde{\psi}$ in discretized space and time variables.

Define $\bar{\underline{\theta}}(t)$ by removing $\theta_7(t)$ from $\tilde{\underline{\theta}}(t)$, and denote the corresponding state space by $\bar{\underline{\boldsymbol{\Theta}}}$. Divide the observation region A into m parts $A_1,\ldots,A_m$, and let $\underline{r}_j$ be a point in A_j, $j = 1,\ldots,m$. Also divide the time interval $(0,\infty)$ into subintervals $(k/n, (k+1)/n]$, $k = 0,1,2,\ldots$, with length $1/n$. Define a step-function approximation $\tilde{z}_{mn}$ of $\tilde{z}$ by

$$\tilde{z}_{mn}(\underline{r},t,\bar{\underline{\eta}}) = \tilde{z}(\underline{r}_j,k/n,\bar{\underline{\eta}}), \quad \underline{r}_j \in A_j,\ j = 1,\ldots,m,$$

$$k/n < t \le (k+1)/n,\ k = 0,1,2,\ldots,$$

and a discretized data $\Delta_{mn}Y$ by

$$\Delta_{mn}Y(j,k) = Y(A_j \times (k/n,(k+1)/n]).$$

Then define a "sampling approximation" $\tilde{\psi}_{mn}$ for almost every realization $(Y,\bar{\underline{\eta}})$ of $(W,\bar{\underline{\theta}}(t))$ by

$$\tilde{\psi}_{mn}(A \times [0],Y,\bar{\underline{\eta}}) = 1,$$

$$\tilde{\psi}_{mn}(A \times [0,(k+1)/n],Y,\bar{\underline{\eta}})$$

$$= \int_{\bar{\underline{\boldsymbol{\Theta}}}} \tilde{\psi}_{mn}(A \times [0,k/n],Y,\underline{\xi})\exp \sum_{j=1}^{m} [\tilde{z}(\underline{r}_j,k/n,\underline{\xi})\Delta_{mn}Y(j,k)$$

$$- \frac{1}{2}\tilde{z}^2(\underline{r}_j,k/n,\underline{\xi})\mu(A_j)/n]P(\bar{\underline{\theta}}(0) \in d\underline{\xi}|\bar{\underline{\theta}}(1/n) = \bar{\underline{\eta}}), \quad k = 0,1,2,\ldots, \tag{6}$$

where $\mu(A_j)$ is the volume of A_j. It is shown that if $z(\cdot,\cdot,\underline{\theta}(\cdot))$ is continuous almost everywhere on $A \times [0,t]$ with probability 1 then, as $m,n \to \infty$ and $\max_{1 \le j \le m} \mu(A_j) \to 0$, $\tilde{\psi}_{mn}$ converges in the stochastic mean to $\tilde{\psi}$.

Observe that (6) is a formula for recursively generating a spatially and temporally sampled version $\tilde{\psi}_{mn}$ of $\tilde{\psi}$. (6) states that with the observation of new data $\Delta_{mn}Y(j,k)$, $j = 1,\ldots,m$, the new value of $\tilde{\psi}_{mn}$ is calculated by weighting the previous value with an exponential function of the new data and by conditionally averaging the weighted value over all possible parameter values. Thus, starting with $\tilde{\psi}_{mn} = 1$ at $k = 0$, we can successively generate $\tilde{\psi}_{mn}$ for each k as a function of the observed data up to the k-th sample. Note the conditional averaging is the same at each k. This is because $\{\underline{\overline{\Theta}}(t),\ 0 \leq t < \infty\}$ is a stationary process. With $\tilde{\psi}_{mn}(A \times [0,k/n],Y,\underline{n})$, $k = 0,1,2,\ldots$, thus generated, the sampled versions of the likelihood ratios and the estimators are immediately obtained by making the appropriate change of symbols in their formulas (2) - (5).

ACKNOWLEDGEMENT

We thank V. E. Benes and E. Wong for their enormous technical assistance.

REFERENCES

0. Kadota, T. T., "Optimal, causal, simultaneous detection and estimation of random signal fields in Gaussian noise field", submitted to the IEEE Trans. on Information Theory.

1. Stratonovich, R. L., Sosulin, Yu, G., "Optimal detection of a Markov process in noise", Eng. Cybern., 6, pp. 7-19, Nov. 1964.

2. Kushner, H. J., "Dynamical equations for optimal nonlinear filtering", J. Diff. Equ., 3, pp. 179-190, April 1967.

3. Lipt'ser, R. S., Shirvaey, A. H., "Nonlinear interpolation of components of Markov diffusion processes", Theory Prob. Appl., 13, 4, pp. 564-583, 1968 (English trans.).

4. Kallianpur, G., Striebel, C., "Estimation of stochastic systems: arbitrary system process with additive white observation errors", Ann. Math. Stat., 39, pp. 785-801, 1969.

5. Wong, E., Stochastic Processes in Information and Dynamical Systems, McGraw-Hill, New York, 1971, pp. 231-241.

6. ________, "A likelihood ratio formula for two-dimensional random fields", IEEE Trans., IT-20, 4, pp. 418-422, 1974.

7. ________, "Recursive filtering for two-dimensional random fields", IEEE Trans., IT-21, 1, pp. 84-86, 1975.

8. Kailath, T., Geesey, R. A., Gevers, M., "An innovations approach to least-squares estimation - IV, V, VI", IEEE Trans., AC-16, pp. 720-727, Dec. 1971; AC-18, pp. 435-453, Oct. 1973: AC-18, pp. 588-600, Dec. 1973.

9. Snyder, D. L., Vaca, M., "A measure transformation approach to estimation and decision for observation derived from martingales", Biomedical Comp. Lab. Monograph No. 271, Washington Univ., St. Louis, Mo., May 1975.

10. Middleton, D., Esposito, R., "Simultaneous optimum detection and estimation of signals in noise", IEEE Trans., IT-14, 3, pp. 434-444, 1968.

11. ____________________________, "New results in the theory of simultaneously optimum detection and estimation of signals in noise", Tech. Memo., T-838, Raytheon Co.

12. Fredriksen, A., Middleton, D., Vandelinde, D., "Simultaneous signal detection and estimation under multiple hypotheses", IEEE Trans., IT-18, 5, Sept. 1972, pp. 607-614.

13. Nahi, N. E., "Optimum recursive estimation with uncertain observation", IEEE Trans., IT-15, 4, pp. 457-462, 1969.

14. Jaffer, A. G., Gupta, S. C., "Coupled detection-estimation of Gaussian processes in Gaussian noise", IEEE Trans., IT-18, 1, Jan. 1972, pp. 106-110.

15. Arques, P. -Y., Macchi, O., Vezzosi, G., "Trois Procedures de detection-estimation simultanees d'un signal", Ann. Telecomm., 28, 11-12, Nov-Dec 1973, pp. 459-468.

16. Lainiotis, D. G., "On a general relationship between estimation, detection, and the Bhattacharyya coefficient", IEEE Trans., IT-15, 4, pp. 504-505, 1969.

17. ________________, "Joint detection, estimation and systems identification", Inform. Control, 19, 1, pp. 75-92, 1971.

18. Park, S. K., Lainiotis, D. G., "Joint detection-estimation of Gaussian signals in white Gaussian noise", Inform. Sci., 4, pp. 315-325, 1972.

THE RECONSTRUCTION OF RANDOM SIGNALS FROM IRREGULARLY SPACED SAMPLES

FREDERICK J. BEUTLER
Computer, Information and Control Engineering Program
The University of Michigan
Ann Arbor, Michigan 48109

ABSTRACT

The following aspects of reconstitution of wide-sense stationary processes from irregularly spaced samples are treated. First, if error-free recovery is possible, recovery is effected by a linear filter which is in a certain sense unique. Second, error-free recovery is related to closedness of certain exponential families in a Banach space; this suggests sampling sequences for which recovery is feasible. Finally, some simple easily implemented interpolatory schemes yield excellent mean square error performance.

I. INTRODUCTION

As every information theorist knows, a signal x(t) without frequency components beyond πW radians per second can be represented by samples $x_n = x(nW^{-1})$ taken W^{-1} seconds apart. Indeed, x(t) can be written as a linear combination of the x_n via the cardinal series. The latter fact implies that it suffices to transmit the samples x_n in place of the entire signal x(t). At the same time, analysis is simplified since the discrete parameter process $\{x_n\}$ contains the same information as x(t).

Various effects limit the validity of the supposition that a truly band-limited signal is regularly sampled without the introduction of noise errors. Here we examine only the consequences of irregular spacing of the samples. Such irregularity may arise from random phenomena such as jitter or inadvertent loss of samples. Alternatively, sampling can occur in some time-shared context, or be governed by external events.

Results on irregular sampling are too extensive (see e.g. [B-1] [B-2] [B-4] [B-5] [B-7] [L-1] [L-2] [L-3] [Y-1] [Y-2]) to permit a survey on this occasion. Accordingly, we can only touch on several relevant questions. First, we give a very general criterion for the recovery of the signal x(t) from the sample sequence $\{x_n\}$; it will follow that if recovery is possible at all, x(t) is given by a liner combination of the x_n (i.e., by a linear filter). The remainder of the paper deals more specifically with band-limited wide-sense stationary x(t) subjected to sampling sequences $\{t_n\}$. Error-free recovery is related to properties of non-harmonic Fourier series, which in turn leads to some sufficiency conditions for error-free recovery. Unfortunately, these sampling schemes do not lend themselves readily to practical realizations, so that we are also motivated to consider very simple signal recovery techniques. Perhaps surprisingly, the latter perform very well--in some instances better than "optimal" filters.

II. LINEARITY OF RECOVERY METHODS

Sampling, as described above, can be regarded as an operator taking each time function $x(t)$ into a sequence $\{x(t_n)\}$. For non-random functions and sampling sequences $\{t_n\}$, we take $\mathscr{S}$ to be the linear space of functions under consideration. The sampling operator L is then linear from $\mathscr{S}$ onto its range $\mathscr{R}(L)$, and all the algebraic properties of linear operators become applicable.

In particular, L has a left inverse (and hence an ordinary inverse) if L is one-one ([N-1], Section 2.7). Moreover, the one-one condition is equivalent to the following: L is one-one iff $x(t_n)=0$ for all n implies that $x(\cdot)$ is the null function of $\mathscr{S}$([N-1], Section 4.4). It is also known that the inverse of a linear operator is again linear. Thus we have the following

Theorem 2.1: For the signals in $\mathscr{S}$, recovery is possible iff $L[x(\cdot)] = \theta$ (the null sequence) implies $x(\cdot) = \theta$ (the null element of $\mathscr{S}$). The recovery filter is then uniquely defined and linear.

Two remarks may be of interest. The above applies not only to simple sampling, but also to simultaneous measurements of $x(t)$ and its derivatives, or to linear functionals on $x(\cdot)$. Second, the uniqueness holds only on $\mathscr{R}(L)$; it we take the range space to be $\mathscr{Y}$, $\mathscr{Y} \supset \mathscr{R}(L)$, $\mathscr{Y} \neq \mathscr{R}(L)$, only the left inverse exists, and this is defined arbitrarily on $\mathscr{Y} \ominus \mathscr{R}(L)$.

Since our principal concerns relate to random processes and perhaps randomly timed sampling, one is naturally interested in an extension of the deterministic case to the stochastic situation. $\mathscr{S}$ now becomes a linear family of stochastic processes. We also suppose $\{t_n\}$ to be a known but possibly random sampling sequence which is independent of each process $x_\alpha(\cdot)$, $\alpha \in \mathscr{S}$. Almost sure error-free recovery over $\mathscr{S}$ can then be accomplished iff

$$P\left[\{x_\alpha(t_n) = 0 \ \forall n\} \cap \{x_\alpha(t) \not\equiv 0\}\right] = 0, \text{ each } \alpha \in \mathscr{S}. \tag{2.1}$$

As before, the recovery filter (which depends on $\{t_n\}$) must be linear on the range of the sampling operator, which is itself random. The variable range does not cause any real difficulty, since algebraic complementation ([N-1], Section 4.12) allows us to enlarge the domain of the recovery operator without destroying its linearity.

An important linear class $\mathscr{S}$ of random processes is the family consisting of all separable wide-sense stationary processes whose spectra have a common support, say $[-\pi, \pi]$. If the sampling instants t_n are almost surely such that $x(t_n) = 0$ (all n) implies $x(t) = 0$ (almost surely, each t) for all processes in $\mathscr{S}$, then the recovery of $x(t)$ from $\{x(t_n)\}$ is feasible, and is accomplished by a linear operation on $\{x(t_n)\}$. In the next Section, we investigate in some detail sequences $\{t_n\}$ which have the above property, and hence permit recovery of $x(t)$.

III. RECOVERY OF BAND-LIMITED PROCESSES

In this Section, we study sampling sequences $\{t_n\}$ that enable us to recover a band-limited wide-sense stationary random processes from their samples $\{x(t_n')\}$. Such processes are described in terms of their spectrum $F(\cdot)$, which we will assume to be supported on $[-\pi, \pi]$. Although these random processes satisfy the conditions of the preceding Section, the additional structure leads to sharper characterization of the recovery problem, and to more explicit results.

Basic to our arguments is the isomorphism derived from the spectral representation of x(t) ([B-2], Section II). Under the isomorphism, the random variable x(t) (in the Hilbert space generated by the random process) corresponds to $e^{i\omega t}$ in a function space. According to the definition of the norms in the Hilbert spaces, we obtain

$$E[\,|x(t) - \sum_1^N c_n^N x(t_n)|^2] = \int_{-\pi}^{+\pi} |e^{i\omega t} - \sum_1^N c_n^N e^{i\omega t_n}|^2 \, dF(\omega) \qquad (3.1)$$

for the mean square difference between x(t) and a linear combination of the samples $x(t_n)$.

In view of (3.1) error-free recovery is equivalent to the convergence to zero of the right side integral. Thus, our recovery problem reduces to finding non-harmonic exponential sets $\{e^{i\omega t_n}\}$ whose linear combinations approximate $e^{i\omega t}$ (any t). In particular, if for $\{t_n\}$ there exist coefficients c_n^N such that

$$\sup_{\omega \in [-\pi, \pi]} |e^{i\omega t} - \sum_1^N c_n^N e^{i\omega t_n}| \to 0, \qquad (3.2)$$

the mean square recovery error goes to zero for any process band-limited to $[-\pi, \pi]$. Moreover, the coefficients c_n^N do not depend on the process or its spectrum F, and (for specified to and $\{t_n\}$) the residual error is bounded by a constant multiplied by $E[\,|x(t)|^2]$.

The class of continuous functions with the sup norm indicated by (3.2) constitutes the Banach space usually denoted by C. Its conjugate (of bounded linear functionals on C) is equivalent to V, the space of functions of bounded variation which are zero at the left end-point $-\pi$. We shall make use of the conjugate space in studying (3.2). To this end, we cite some well known results concerning Banach spaces.

Let X be a Banach space having X^* as its conjugate. A set $\{x_k\} \in X$ is said to be closed in X if

$$x^*(x_k) = 0 \quad \text{all } k, \; x^* \in X^* \qquad (3.3)$$

implies x^* is the null element of X^*. But closure is equivalent to completeness ([D-1], p. 263), viz.

Lemma 3.1: Let $\{x_k\}$ be closed in X. Then for any $x \in X$ there exist coefficients $\{c_n^N\}$ such that

$$\|\, x - \sum_1^N c_n^N x_k \,\| \to 0 \qquad (3.4)$$

in X norm.

Our application of Lemma 3.1 is to the set $\{\exp(i\omega t_n)\}$ in C. In fact, (3.4)--and hence (3.2)--holds if for $G \in V$

$$\int_{-\pi}^{+\pi} [\exp(i\omega t_n)] \, dG(\omega) = 0, \text{ all } n \qquad (3.5)$$

we have $G = \theta$. It is seen that the criterion (3.5) is similar in form to the condition established in Section II; however, (3.5) is easier to verify. Indeed, the literature contains classes of non-harmonic exponentials for which (3.5) is true.

To obtain some examples of $\{t_n\}$ admitting error free recovery, we introduce the notation $\Lambda^+(u)$ for the number of t_n in the interval $(0, u)$. Use of a theorem by Levinson ([B-4]; [L-3], Corollary 1) then shows that error free recovery is possible provided

$$\overline{\lim_{u \to \infty}} \; [\Lambda^+(u)/u] > 1. \tag{3.6}$$

This means that any one-sided sequence of t_n which occasionally exceeds the Nyquist rate suffices to assure recovery of any wide-sense stationary process whose spectrum is concentrated on $[-\pi, \pi]$.

By supposing the sampling sequence to be $\{-t_n\}$ in place of $\{t_n\}$, $x(t)$ can be recovered from earlier samples extending to $-\infty$. A translation by t can then be used to prove that $x(t)$ can always be recovered by a linear causal operation on the samples of the past. It is also clear that the recoverability of $x(t)$ is unaffected by the deletion of any finite number of samples, or even an infinite number if the deletions are asymptotically sparse; this is a consequence of the observation that the lim sup in (3.6) remains unchanged by the deletions.

The deterministic sampling sequence $\{t_n\}$ may be replaced by a stationary point process [B-6] under appropriate conditions; more precisely, we hypothesise that $\{t_n\}$ is ergodic, with an average number of points per unit interval exceeding unity. The class of such point processes includes the Poisson, periodic subject to skips and/or jitter, self-generating burst processes, as well as superpositions of these. In each case, (3.6) is satisfied with probability one.

Other closure theorems suggest further suitable sampling sequences [B-4]. For example, $t_n = n - \sqrt{n}$ satisfies (3.5) but not (3.6). In fact, there are sequences at less than the Nyquist rate which permit recovery of the signal. One supposes that the latter are unstable, in the sense that a small change in the samples $x(t_n)$ leads to a large variation in the reconstituted $x(t)$.

IV. PERFORMANCE OF SIMPLE INTERPOLATORS

For practical sampling applications, it is not enough to show that error-free recovery is theoretically possible. In general, the recovery filter cannot be specified; even if this filter were known, its susceptibility to noise may preclude its use. We may therefore wish to consider some simple filters utilizing only a few samples. The error analysis of such interpolation (and extrapolation) filters constitutes the objective of this Section.

We shall suppose a real wide-sense stationary $x(t)$, which is standardized to $E\{[x(t)]^2\} = 1$. Sampling is accomplished in accordance with a stationary point process $\{t_n\}$ [B-6] independent of $x(t)$; as was noted previously, stationary point processes represent the types of sampling which may be realistically encountered.

One is interested principally in relatively small mean-square recovery errors, which requires a high sampling rate. Here "high" must be interpreted in dimensionless units. If W is some measure of the bandwidth of x(t), and β is the mean sampling rate, we define $\gamma = W^{-1}\beta$ as the non-dimensional sampling rate. Our results will then appear as asymptotic formulas valid for large γ. This philosophy also has the advantage of yielding expression easy to evaluate and compare.

Let us now calculate the (asymptotic) mean square error when the recovery is based on the single sample immediately preceding the present time t. We consider in particular the interpolator whose output is

$$\hat{x}(t) = x(t - L_{-1}(t)) \exp[-bL_{-1}(t))] , \tag{4.1}$$

in which $b \geq 0$ is the exponential decay constant, and $L_{-1}(t)$ is the time interval between t and the sampling time just before t. In general, the mean square recovery error is

$$\epsilon^2 = E\{[x(t) - \hat{x}(t))]^2\} = 1 - 2E[x(t)\hat{x}(t)] + E\{[\hat{x}(t)]^2\}, \tag{4.2}$$

in which the expectation is taken successively on x(t) and $\{t_n\}$. When the correlation R_x of x(t) is written as the inverse transform of the spectral density S_x, we obtain for the mean square error

$$\epsilon^2 \doteq + \tilde{g}_1^*(2\alpha\gamma^{-1}) - \pi^{-1} \int_{-\infty}^{\infty} S_{\tilde{x}}(\omega)\tilde{g}_1^*(\alpha\gamma^{-1} - i\omega\gamma^{-1})\, d\omega . \tag{4.3}$$

In (4.3), the notation is as follows: $\alpha = b/W$ is the normalized (dimensionless) decay parameter, g_1^* is the generating function (i.e., Laplace transform) of L_{-1} and hence also of the forward recurrence time L_1 (the two have identical statistics--see [B-6]), and the tildes denote normalized processes. Specifically, $\{\tilde{t}_n\}$ is normalized to unit average rate, and $\tilde{x}(t)$ has unit bandwidth (by whatever bandwidth measure has been chosen). Now if L_1 is of finite mean square, g_1^* is expandable as a three term Maclaurin series with remainder. When this series is substituted into (4.3), we obtain the asymptotic expression [B-4]

$$\epsilon^2 = \gamma^{-2}(\alpha^2 + E\{[\dot{\tilde{x}}(0)]^2\})E\{[\tilde{L}_1(0)]^2\} + o(\gamma^{-2}). \tag{4.4}$$

From (4.4) we may conclude that (1) the mean square error varies as γ^{-2}, (2) sample-and-hold interpolation (i.e., $\alpha = 0$) is preferable to any decay, (3) the error increases with the irregularity of the sampling sequence.

Of course, (4.4) is applicable only to differentiable signals. When $R_x(\tau) = \exp[-a|\tau|]$, a somewhat different computation (starting again with (4.2))yields

$$\epsilon^2 = 2\gamma^{-1} E[\tilde{L}_1(0)] + o(\gamma^{-1}) \tag{4.5}$$

if we take W=a (half power point). It is of interest to compare (4.5) with the error incurred for Wiener optimum extrapolation. For instance, if $\{t_n\}$ is a Poisson process, $E[\tilde{L}_1(0)] = 1$ in (4.5), and the Wiener extrapolation error is [B-3]

$$\epsilon^2 = \sqrt{2}\gamma^{-\frac{1}{2}} + o(\gamma^{-1}) . \tag{4.6}$$

It seems paradoxical that the "optimum" extrapolator gives a larger error than a rather crude ad-hoc device; however, it is recalled that Wiener theory selects the optimum time-invariant filter, while the sample-and-hold procedure represents a time varying operation.

The mean square error corresponding to two-point polygonal (straight-line) interpolation and extrapolation can also be computed. Here we have for interpolation

$$x(t) = \frac{L_{-1}\, x(t-L_1) + L_1 x(t-L_{-1})}{L_1 + L_{-1}} \tag{4.7}$$

in which the arguments of L_1 and L_{-1} are suppressed for convenience. Asymptotic approximations on (4.2) involve a series expression for R_x valid if $x(t)$ is twice differentiable. There results [B-4]

$$\epsilon^2 = \tfrac{1}{4} \gamma^{-4} E\,\{[\ddot{\tilde{x}}(0)]^2\}\, E\,\{[\tilde{L}_1(0)\,\tilde{L}_{-1}(0)]^2\} + o(\gamma^{-4}). \tag{4.8}$$

The causal extrapolator is subject to a similar formula, but with a larger coefficient multiplying the γ^{-4}. For non-differentiable signals the calculation of ϵ^2 appears insuperably difficult, although some special cases seem to indicate that $\epsilon^2 = 0(\gamma^{-1})$. We are again able to make some comparisons with Wiener filters utilizing the entire past (rather than just two points); we find that $\epsilon^2 = 0(\gamma^{-1})$, which contrasts unfavorably with the $0(\gamma^{-4})$ form for the two-point straight line extrapolator.

Sample-and-hold and polygonal interpolation are merely special cases of polynomial interpolators, in which the n-1 order polynomial P_{n-1} approximates a function f by passing through the n points $f(t_k)$, $k=1, 2, \dots, n$. The error or Cauchy remainder ([D-1] , Section 3.1) is

$$|f(t) - P_{n-1}(t)| = \frac{|f^{(n)}(s)|}{n!} \prod_1^n |t-t_k| \ , \tag{4.9}$$

where s is unspecified other than $\min(t_k) \leq s \leq \max(t_k)$. This property of s makes it impossible to evaluate the mean square value (error) in the customary manner, since s depends stochastically both on $x(t)$ and on $\{t_n\}$ when (4.9) is applied to a randomly sampled random signal.

Fortunately, the difficulty just alluded to disappears when we normalize as before, and let $\gamma \to \infty$. In heuristic terms, the effect of these operations is to "squeeze" the interval down to a point. Accordingly, (4.9) ultimately leads to

$$\epsilon^2 = \gamma^{-2n} (n!)^{-2} E\{[\tilde{x}^{(n)}(0)]^2\}\, E \left(\prod_1^n [\tilde{L}_{jk}(0)]^2 \right) + o(\gamma^{-2n}), \tag{4.10}$$

for a n times continuously differentiable $x(t)$. In (4.10), the indices j_k are given by $L_{j_k}(t) = t_k - t$.

We sum up: For an n times differentiable signal, the mean square recovery error varies as γ^{-2n} at high sampling rates, with multiplicative factors reflecting the shape of the spectrum of $x(t)$, and the irregularity of the sampling sequence $\{t_n\}$; this result holds for both causal and noncausal recovery schemes, but with different multiplicative factors. It may be added that such deviations from periodicity of $\{t_n\}$ as jitter, random skipping--even for Poisson $\{t_n\}$--mean square errors are not greatly enlarged.

It is conjectured that a non-differentiable signal generates a mean square recovery error asymptotically proportional to γ^{-1}; this is borne out by the computations it is feasible to make. While the performance for non-differentiable signals is not as satisfactory as that for differentiable signals, it is nevertheless an improvement over the results obtained by a time-invariant Wiener optimum recovery filter.

ACKNOWLEDGEMENT

The author is pleased to acknowledge the support of the National Science Foundation under NSF Grant No. ENG 75-20223 in the preparation of this manuscript.

REFERENCES

[B-1] A. V. Balakrishnan, "On the problem of time jitter in sampling," IRE Trans. Info. Th., IT-8 (1962), 226-236.

[B-2] F.J. Beutler, "Sampling theorems and bases in a Hilbert space," Information and Control, 4 (1961), 97-117.

[B-3] F.J. Beutler, "Prediction for wide-sense Markov processes," IRE Trans. Info. Th., IT-7 (1961), 267-269.

[B-4] F.J. Beutler, "Error-free recovery of signals from irregularly space samples," SIAM Rev., 8 (1966), 328-335.

[B-5] F.J. Beutler, "Recovery of randomly samples signals by simple interpolators, Information and Control, 26 (1974), 312-340.

[B-6] F.J. Beutler and O. Leneman, "The theory of stationary point processes," Acta Math., 116 (1966), 159-197.

[B-7] W.M. Brown, "Sampling with random jitter," J. SIAM, 11 (1961), 460-473.

[D-1] P.J. Davis, Interpolation and Approximation, Blaisdell, New York, 1963.

[L-1] O. Leneman, "Random sampling of random processes: optimum linear interpolation," J. Franklin Inst., 281 (1966), 302-314.

[L-2] O. Leneman and J. Lewis, "Random sampling of random processes: mean-square comparison of various interpolators," IEEE Trans. Auto. Control, 11 (1966), 394-403.

[L-3] B. Liu and T. Stanley, "Error bounds for jittered sampling," IEEE Trans. Auto. Control, 10 (1965), 449-454.

[L-4] N. Levinson, Gap and Density Theorems, Colloq. Pub. 26, Am. Math. Soc., Providence, R.I., 1940.

[N-1] A. Naylor and G. Sell, Linear Operator Theory in Engineering and Science, Holt, Rinehart and Winston, New York, 1971.

[Y-1] K. Yao, "Application of reproducing kernel Hilbert spaces -- band-limited signal models," Information and Control, 11 (1967), 429-444.

[Y-2] K. Yao and J. Thomas, "On some stability and interpolatory properties of nonuniform sampling expansions," IEEE Trans. Circuit Th., CT-14 (1967), 404-408.

ANALYSIS OF PSEUDO-RANDOM NUMBER SEQUENCES GENERATED BY A METHOD OF RADER, RABINER, AND SCHAFER*

TING-PIEN FANG and CHARLES E. MOLNAR
Computer Systems Laboratory
Washington University
St. Louis, Missouri 63110

ABSTRACT

We have studied algebraic properties of RRS pseudo-random number sequences. An M-bit random number $W_M(n)$ is generated from its two precedents as: $W_M(n)=P_M[W_M(n-1) \oplus W_M(n-2)]$, where $P_M[.]$ denotes a cyclic M-bit permutation and $\oplus$ denotes bit-wise exclusive-or. Each RRS sequence is determined by its two initial values $W_M(0)$ and $W_M(1)$. We have found: 1. a relation of $W_M(n)$ to $W_M(0)$ and $W_M(1)$, 2. best choices of word length, cyclic permutation, and initialization for a long sequence, and 3. properties shared by a class of RRS sequences with certain word lengths, such as 3, 5, 11, 13, 19, 29, 37, etc. .

I. INTRODUCTION

Most random number generators are either multiplicative or additive [1]. A multiplicative generator is slow because multiplication is used. An additive generator needs at least 6 memory words for storing previous numbers in order to generate good random number sequences [2]. In this paper, we have studied an algorithm for generating random number sequences devised by Rader, Rabiner, and Schafer [3]. This algorithm is easy to implement, needs only two words of memory, and can provide a new number in a very short time. We shall call this algorithm the RRS algorithm; sequences generated by this algorithm, RRS sequences; and numbers from an RRS sequence, RRS numbers. We shall also use the term RRS word when we wish to emphasize the values of individual bits in an RRS number, rather than its integer or fractional value.

---

* This work has been supported by the Advanced Research Projects Agency under Contract SO-302, by the Division of Biotechnology Resources Program of the National Institutes of Health under Grant RR-00396 and by the Division of Health Services Research and Development of the Public Health Service under Grant 5-R18-HS-01533.

The RRS algorithm for generating an M-bit random number sequence is

$$W_M(n)=P_M[W_M(n-1) \oplus W_M(n-2)], \qquad (1)$$

where P_M [.] denotes a cyclic M-bit permutation*;

$\oplus$ denotes a mod 2 addition (i.e. bit-wise exclusive-or);

and $W_M(0)$ and $W_M(1)$ are two M-bit initial words.

We shall subsequently omit the suffix M except when it is needed for clarity; likewise we will generally use the term sequence to mean RRS sequence.

To construct an RRS sequence generator, one has three choices to make: 1) Word length M; 2) Initial words W(0) and W(1); and 3) Cyclic permutation.

A hardware implementation of an RRS algorithm, as shown in Figure 1, needs only two M-bit registers to store the words W(n-1) and W(n-2) and M exclusive-or gates to perform the mod 2 additions. The particular permutation P determines the connections of the exclusive or outputs to the lower row of flip-flops. The minimum operation time is equal to the sum of one flip-flop propagation delay and one exclusive-or gate propagation delay. An RRS generator using ECL logic can generate RRS numbers at a rate higher than 10^8 words/sec..

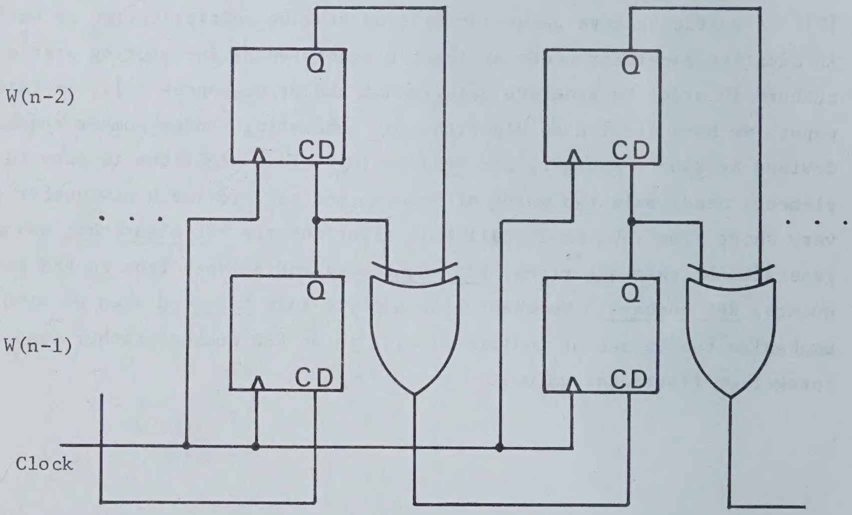


Figure 1. A hardware RRS generator

---

* This is a slight generalization of the RRS algorithm as originally described, which specified a cyclic rotation by a number of steps prime to the word length rather than a general cyclic permutation.

To be useful, a pseudo-random number sequence must:

1) have a long enough period; and
2) have properties consistent with a suitable random process model.

In this paper, we have studied algebraic properties of RRS sequences. The value of this algebraic analysis lies in its ability to determine the period of a specific sequence and to call attention to certain regularities in sequences that are not obvious and may pose difficulties in some kinds of applications.

The algebraic analysis alone does not constitute a sufficient set of tests for the adequacy of RRS numbers for most applications, and additional statistical tests of a traditional kind are still required. Even if a sequence appears to be random with respect to a particular set of tests, it may not appear to be random with respect to some other test. Given knowledge of the generating algorithm, one can always devise a test that the sequence will fail. In this sense there is no universally applicable pseudo-random number sequences.

As will be discussed in detail later (Section III and IV), the period N of a sequence is dependent upon the choice of M and upon the initial words W(0) and W(1), but not on the particular cyclic permutation P. We define the maximum period of a sequence with a particular choice of M as the longest period that can occur with any initialization.

Table 2 in the paper by Rader et.al. [3] gives maximum periods for RRS sequences of word lengths up to 25. It is apparent that while maximum periods tend to increase as M increases, there is much irregularity in this trend and nearby values of M may produce widely different periods.

We define long RRS sequences as sequences with periods that are equal to or greater than $3(2^{M-1}+1)$. We have proved [4] that a necessary condition on M for generation of a long RRS sequence is:

Condition 1

The smallest positive integer h for which $2^h \bmod M=1$ is $M-1$.

The only values of M less than 53 which satisfy Condition 1 are 3, 5, 11, 13, 19, 29, and 37, and we have found that the corresponding RRS sequences have long maximum periods in each of these cases. It is our unproved conjecture that Condition 1 is sufficient as well as necessary for a long maximum period. We shall identify values of M for which Condition 1 is satisfied as M_L.

Another useful classification of RRS sequences can be made according to the values of the initial words W(0) and W(1). We define three such classes, all of which have initial word W(0)=00...0: 1) An Even RRS sequence has W(1) containing an even and non-zero number of '1's; 2) An Odd RRS sequence has W(1) containing an odd number of '1's and at least one '0'; and 3) A Seed RRS sequence has W(1)=10...0, we note that a Seed sequence is also an Odd sequence.

We shall define one additional class of RRS sequence on the basis of the permutation P. If P is a cyclic rotation of one position to the right, we will call the sequence a Special RRS sequence. This classification is independent of the previous classifications based on word length and initialization, so any other class of RRS sequence contains members that are special RRS sequences and members that are not.

In Sections II and III, we shall discuss a polynomial representation of RRS words, and define a mapping function for RRS words. Polynomial representation of RRS words provides a useful tool for characterizing RRS sequences, and the mapping function of RRS words relates an individual RRS word to the initial words W(0) and W(1) without a need for generating the intermediate words of the sequence. Properties of mapping functions help us in finding candidates of sequence for further analysis. In Section IV, we use a 5-bit Seed Special RRS sequence as an example to illustrate regularities in M_L-bit Seed Special RRS sequences found by algebraic properties and use these algebraic properties to characterize M_L-bit RRS sequences. Finally, in Section V, we show how algebraic results are applied to reduce the length and number of tests needed to verify a specific conjecture.

II. POLYNOMIAL REPRESENTATION OF RRS WORDS

Definition 1: Polynomial representation of RRS words

A polynomial $f_n(x)$ is said to be the polynomial representation of an RRS word $W(n)=a_{n,0};a_{n,1};\ldots;a_{n,M-1}$, where $a_{n,m}$ represents the mth bit of the nth word in the sequence, where m=0,1,...,M-1, if

$$f_n(x) = a_{n,0}x \oplus a_{n,1}x^2 \oplus \ldots \oplus a_{n,M-1},$$

$$= \sum_{m=0}^{M-1} a_{n,m}x^{m+1}, \tag{2}$$

$$\text{where } x^M = x^0 = 1.$$

With the polynomial representation of RRS words, it becomes easier for us to characterize RRS sequences. All sequences discussed hereafter in this Section are Seed Special RRS sequences, unless otherwise specified. The term Seed is chosen to specify this class of sequences for reasons that will become clear later.

The polynomial representation version of Equation (1) is:

$$f_n(x) = x[f_{n-1}(x) \oplus f_{n-2}(x)]. \tag{3}$$

Since the additions are mod 2, it is easy to prove that:

1. $$f_n^2(x) = \sum_{m=0}^{M-1} a_{n,m} x^{2(m+1) \bmod M}, \text{ and} \qquad (4)$$

2. $$[f_i(x) \oplus f_j(x)]^2 = f_i^2(x) \oplus f_j^2(x). \qquad (5)$$

Many properties of sequences can be found with the help of polynomial representation of RRS words, and they are given in the form of theorems. For theorems for which proofs are straightforward, we provide equation number(s) and/or theorem number(s) that may be used to prove them.

Theorem 1.

$$f_{2i \bmod N}(x) = f_i^2(x), \text{ where } i=0,1,2,\ldots,N-1. \qquad (6)$$

[Use Equations (3) and (4)].

Corollary 1.

$$f_{2^i \bmod N}(x) = x^{2^i \bmod M}, \text{ where } i=0,1,2,\ldots\ . \qquad (7)$$

Theorem 2. (Offset Theorem)

$$f_{(i+2j) \bmod N}(x) = [f_i(x) \cdot x^j] \oplus [f_j(x) \cdot f_{(i+j) \bmod N}(x)], \qquad (8)$$

where i,j=0,1,2,...,N-1.

[Use Equation (3)].

Corollary 2

$$f_{(2^i+1) \bmod N}(x) = x^{(2^i+1) \bmod M}(1 \oplus \sum_{j=0}^{i-1} x^{-2^j \bmod M}), \qquad (9)$$

where i=0,1,2,... .

[Use Theorem 2 and Corollary 1].

III. MAPPING FUNCTIONS

We have proved [4] that given a sequence, R, there exists corresponding to each bit $a_{n,m}$, a function $M^P(n,m)=Q^P(n,m);R^P(n,m)$ which relates $a_{n,m}$ directly to the initial words W(0) and W(1) of R according to the following equation:

$$a_{n,m} = [Q^P(n,m) \cdot W(0)] \oplus [R^P(n,m) \cdot W(1)], \qquad (10)$$

where n=0,1,2,...,N-1,m=0,1,2,...,M-1, and "." denotes a dot product such that for $A=a_0;a_1;\ldots;a_{M-1}$ and $B=b_0;b_1;\ldots;b_{M-1}$, $A \cdot B = a_0 \cdot b_0 \oplus a_1 \cdot b_1 \oplus \ldots \oplus a_{M-1} \cdot b_{M-1}$.

We call $M^P(n,m)$ the mapping function of $a_{n,m}$.

We have found [4] several properties of mapping functions as listed below:

1. $Q^P(n,m)$ [or $R^P(n,m)$] forms an RRS sequence with cyclic permutation P^{-1} (inverse of P) and initial words $Q^P(0,m)$ and $Q^P(1,m)$ [or $R^P(0,m)$ and $R^P(1,m)$];
2. Relations between $Q^P(n,m)$ and $R^P(n,m)$:
$$Q^P(n+1,m) = P^{-1}[R^P(n,m)], \text{ and} \qquad (11a)$$
$$R^P(n+1,m) = Q^P(n,m) \oplus Q^P(n+1,m); \qquad (11b)$$
3. If $a_{n,i} = P^k a_{n,j}$, where k=0,1,2,..., then $Q^P(n,i)=P^k[Q^P(n,j)]$, and $R^P(n,i)=P^k[R^P(n,j)]$; and
4. The mapping function sequences of all M-bit RRS sequences have the same period independent of the choice of cyclic permutation.

A table of two 3-bit RRS sequences generated by two different cyclic permutations is given below. Included in this table are several mapping functions which will help to illustrate above stated properties.

	sequence 1 P:rotation to the right by one step					sequence 2 P:rotation to the left by one step		
	W(n)	Q(n,0)	R(n,0)	Q(n,1)	R(n,1)	W(n)	Q(n,0)	R(n,0)
0	000	100	000	010	000	000	100	000
1	100	000	100	000	010	100	000	100
2	010	001	001	100	100	001	010	010
3	011	010	011	001	101	011	001	011
4	100	110	100	011	010	100	101	100
5	111	001	111	100	111	111	010	111
6	101	111	110	111	011	110	111	101
7	001	101	010	110	001	010	110	001
8	010	100	001	010	100	001	100	010
9	101	010	110	001	011	110	001	101
10	111	101	111	110	111	111	110	111
11	001	111	010	111	001	010	111	001
12	011	100	011	010	101	011	100	011
13	001	110	010	011	001	010	101	001
14	001	100	010	010	001	010	100	001
	repeats							

We have shown that mapping functions provide a direct way of relating the n-th word to the initial word pair of an RRS sequence. From perperties 1, 2, and 3, we can prove that for a fixed word length M, the maximum possible period of an M-bit RRS sequence is the period of its mapping function sequence. No RRS sequence can have a period longer than the period of its mapping function sequence, because if its mapping function sequence repeats, so does this sequence. In fact the period of an M-bit RRS sequence is an integer factor of the period of its mapping function sequence. Consequently, a maximum period is guaranteed if one of the initial words contains all

'0's and the other one contains only one '1'. Also, from property 4, we know that the maximum period of an M-bit RRS sequence is independent of the choice of cyclic permutation in the generating algorithm. Thus, we have found that Seed Special RRS sequences are good candidates for algebraic analysis. In the next section, we shall use the 5-bit Seed Special sequence to illustrate algebraic properties (regularities) found in M_L-bit Seed Special sequences, and we shall use them to characterize M_L-bit sequences.

IV. CHARACTERISTICS OF M_L-BIT RRS SEQUENCES

Let R_5' denote the 5-bit Seed Special sequence. Several important regularities in R_5' are shown in Table 1:

1. There are exactly 5 words with the value 00000 in R_5', and they are uniformly distributed in R_5', i.e. distances between adjacent 00000's are identical;
2. In addition, between two adjacent 00000's there are two words with the value 11111, and the distance between each 00000 and its closest 11111, and the distance between any adjacent pair of 11111's are all identical and equal to d, as shown in Table 1, and
3. Corresponding words in segments 1, 2, 3, 4, and 5 are permutationally related. Thus, given any segment of R_5', words in the other 4 segments can be obtained by cyclic permutation of their corresponding words in the initial segment. These segments are called minimum sub-sequences (see Appendix for definition).

From these regularities, we know that the period N_5' of R_5' is $3\cdot 5\cdot d$. From the property of M_L, we can prove [4] that these regularities will also occur in all M_L-bit Odd sequences, (except when $M_L=3$, which is a special case, because the minimum sub-sequence property does not exist in any 3-bit RRS sequence. We shall not include 3 in M_L, hereafter in this paper). For example, as there are 5 minimum sub-sequences in R_5', there are M_L minimum sub-sequences in each M_L-bit Odd sequence.

From the property of M_L and Corollary 2, we have proved [4] that in each M_L-bit Odd sequence R_{M_L}, the word $W(2^{M_L-1}+1)$ has the value 11...1. If $W(2^{M_L-1}+1)$ is the first word in R_{M_L}, with the value 11...1, then the period N_{M_L} of R_{M_L} is

$$N_{M_L} = 3\cdot M_L(2^{M_L-1}+1). \qquad (12)$$

Otherwise, N_{M_L} is an integer factor of $3 \cdot M_L(2^{M_L-1}+1)$. This is clearly shown in the example of R'_5. If $W(2^{M_L-1}+1)$ [in this case W(17)] is the first word in R'_5 with the value 11111, then d=17, thus $N'_5=3 \cdot 5 \cdot 17=255$. Otherwise d would be an integer factor of 17, consequently, N'_5 would become an integer factor of $3 \cdot 5 \cdot 17$.

We have found that when M_L = 5, 11, 13, 19, in an M_L-bit Odd sequence R_{M_L}, $W(2^{M_L-1}+1)$ is the first word with the value 11...1, therefore, Equation (12) is true for Odd sequences with word lengths M_L = 5, 11, 13, 19. It is worth mentioning that periods of M_L-bit Even sequences are one-third of those of M_L-bit Odd sequences with the same word lengths and there are M_L minimum sub-sequences leading by 00..0's, in each M_L-bit Even sequence [4]. We have also proved [4] that when M_L=5, 11, 13, 19, due to the fact that N_{M_L} satisfies Equation (12), every M_L-bit RRS sequence contains at least one word with the value 00...0. With this in mind and with the minimum sub-sequence properties stated above, we have derived a quick way of generating all 5, 11, 13, 19-bit RRS sequences, i.e. by fixing W(0) to the value 00...0 and using different values for W(1). Furthermore, when M_L = 5, 11, 13, 19, with a fixed P, every M_L-bit RRS sequence can be generated by the Seed RRS sequence with the same P and M_L. (This is why we called this particular sequence the Seed RRS sequence).

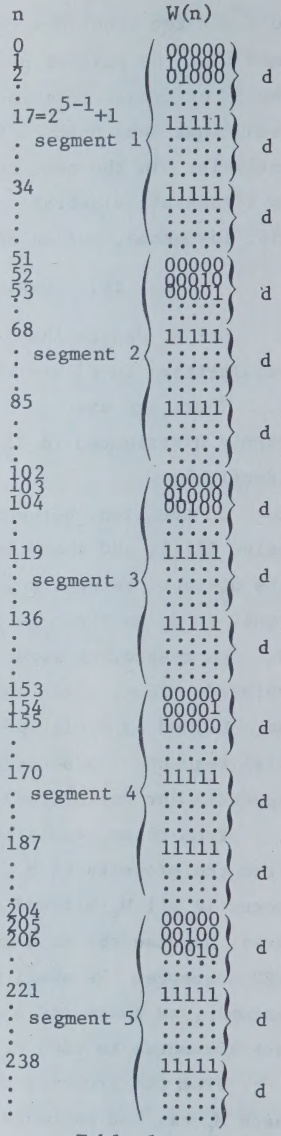

n
W(n)
0
1
2
$17=2^{5-1}+1$
34
51
52
53
68
85
102
103
104
119
136
153
154
155
170
187
204
205
206
221
238
segment 1
segment 2
segment 3
segment 4
segment 5
00000
10000
01000
11111
00000
00010
00001
00000
01000
00100
00000
00001
10000
00000
00100
00010
d

Table 1
Regularities in R'_5

One important regularity in all RRS sequences that we have to mention is the parity of an RRS word (Definition A.3). In some RRS sequences, such as Even RRS sequences, all words have parities '0', while in some other sequences, such as Odd RRS sequences, one third of the words have parities '0' and two-thirds of the words have parities '1', alternating as 011011... with increasing n.

We have shown properties of M_L-bit RRS sequences when N_{M_L} satisfies Equation (12). Although we believe that Equation (12) is true for Odd RRS sequences with other values of M_L (greater M_L's), so far we have not been able to prove it. We have built a hardware system with Restructured Macromodules (RMM) [5,6] to verify that periods of 29-bit and 37-bit Odd RSS sequences (29 and 37 are the next two M_L's greater than 19) also satisfy Equation (12). We have used properties stated in this paper to dramatically reduce the length of test needed in this experiment. This will be discussed in Section V.

V. AN EXAMPLE SHOWING THE APPLICATION OF ALGEBRAIC PROPERTIES IN SIMPLIFYING AN EXPERIMENT

Let us look at the experiment when $M_L=37$, we want to prove that a 37-bit Odd sequence will have a period as given in Equation (12). We have shown in Section IV that an Odd sequence has the same period as that of its Seed sequence. Thus, we only have to prove that the period of the 37-bit Seed sequence satisfies Equation (12). Furthermore, we have proved in Section III that the period of an RRS sequence with fixed M and W(0) and W(1) is independent of the choice of the cyclic permutation used in the generating algorithm, therefore, we only have to prove that the period N'_{37} of the 37-bit Seed Special sequence R'_{37} satisfies Equation (12). We know that N'_{37} is either equal to $3\cdot 37\,(2^{37-1}+1)$ or is an integer factor of $3\cdot 37(2^{37-1}+1)$. We have shown that in R'_{37}, the word $W(2^{36}+1)$ has the value 11...1, and if $N'_{37}<[3\cdot 37(2^{36}+1)]$, then the first word with the value 11...1 must occur before $2^{36}+1$ words have been generated. By decomposing $2^{36}+1$, we find that its smallest factor is 17. Therefore, we know that the first word, say the n-th word, with the value 11...1 occurs either at $n=2^{36}+1$, or $n \leq (2^{36}+1)/17$. Since $[(2^{36}+1)/17]<2^{32}$, we only have to generate and examine 2^{32} words, instead of $3\cdot 37(2^{36}+1)$ which is about 1.7×2^{42}, to determine the period N'_{37}.

Since $3\cdot 37(2^{36}+1)$ is about 1776 times 2^{32}, there is quite a saving of time. The total operation time for generating and examining 2^{32} 37-bit RRS words was 3 hours and 12 minutes. If without the help of properties found in this paper all $3\cdot 37(2^{36}+1)$ 37-bit RRS words were generated and examined to determine the period of R'_{37}, then the operation time would have been

1776·3.2 hours = 236 days, 19 hours, and 12 minutes.

We have found that N'_{37} and N'_{29} (period of the 29-bit Seed Special sequence) satisfy Equation (17). Hence, all properties stated in Section IV are true for 29 and 37 bit sequences.

There were, on the average, 24.5 different macromodule operations in generating and examining each 37-bit RRS word, which means there are a total of $1.05X10^{11}$ (105 billion) operations in 3 hours and 12 minutes. Thus, we have also incidentally tested the reliability of RMM hardware used in this configuration, since the calculation precisely verified a very specific conjecture.

VI. SUMMARY

We have mentioned in the beginning that to use an RRS generator, one has three choices to make:

1. Word length: M;
2. Initial words: W(0) and W(1); and
3. Cyclic permutation: P.

From our studies, we can give some recommendations on these choices:

1. On word length: Word length is suggested by the word length of the machine being used; for a 16-bit machine, a 19-bit RRS sequence is a good candidate, and will yield a period of 14,942,265. For applications needing longer periods, one may use a 29-bit RRS sequence with a period of 23,353,884,759 or a 37-bit RRS sequence with a period of 7,627,861,917,807.
2. On initial words: Initial words of Odd RRS sequences are good initial words, i.e. the longest period for that word length is guaranteed.
3. On cyclic permutation: We have performed several preliminary statistical tests on 19-bit RRS sequences, and results are in general encouraging. Except for run tests [1], results appear insensive to the choice of cyclic permutation. We have found that when rotation to the right by one step or rotation to the left by one step is used, the run lengths are much shorter than predicted by a random process model that assumes independent samples from a uniform distribution, suggesting that these two rotations are bad choices. Although we have not done a thorough study on the choice of permutations, we have found that if rotation to the right by 7 steps is used, this problem disappears. This anomaly shows the need for statistical tests, in addition to analysis.

In this paper, we have also illustrated certain regularities in M_L-bit RRS sequences which we think a user should note.

REFERENCES

1. Knuth, D. E., "The Art of Computer Programming", Volume 2: Seminumerical Algorithms, Addison-Wesley Company, 1971.

2. Green, B. F. Jr., Smith, J. E. K., and Klem, L., "Empirical Tests of an Additive Randon Number Generator", Journal of the Association for Computing Machinery, 6, 527-537, 1959.

3. Rader, C. M., Rabiner, L. R., and Schafer, R. W., "A Fast Method of Generating Digital Random Numbers", The Bell System Technical Journal, Volume 49, No. 9, 2303-2310, November 1970.

4. Fang, T. P., "Analysis of Pseudo-Random Number Sequences Generated by a Method of Rader, Rabiner, and Schafer", Doctoral Dissertation, Sever Institute of Technology, Washington University, St. Louis, Missouri, (to be submitted in December 1976).

5. Clark, W. A., and Molnar, C. E., "Macromodular Computer Systems", Computers in Biomedical Research, Edited by Stacy, R. W., and Waxman, B. D., Volume IV, Chapter 3, Academic Press, New York and London, 1974.

6. "Macromodular Computer Design, Part 3, Restructured Macromodules", Technical Report No. 49, Computer Systems Laboratory, Washington University, St. Louis, Missouri, 1974.

APPENDIX

Definition A.1. Subsequence

A set of i consecutive words, e.g. W(n) through W[(n+i-1) mod N], of an RRS sequence, R is said to be a sub-sequence of R, if for j=0,1,2,..., i-1, each W[(n+i+j) mod N] can be generated by a fixed cyclic permutation of W[(n+j) mod N].

Definition A.1 does not preclude the possibility that there are sub-sequences contained entirely within a given sub-sequence. In order to distinguish sub-sequences that consist of one sub-sequence from sub-sequences that consist of more than one sub-sequence, we introduce the following definition.

Definition A.2. Minimum Sub-Sequence

A sub-sequence of an RRS sequence, R, is said to be a minimum sub-sequence of R, if it consists of only one sub-sequence.

Definition A.3. Parity of an RRS Word

The parity p_n of an RRS word is defined as follows:

$$p_n = a_{n,0} \oplus a_{n,1} \oplus a_{n,2} \oplus \ldots \oplus a_{n,M-1}. \tag{A1}$$

An RRS word is said to have parity '0' if p_n=0, or a parity '1' if p_n=1. From Equation (1) and Definition A.3, we know that

$$p_n = p_{n-1} \oplus p_{n-2}. \tag{A2}$$

It is easy to prove from Equation (A2) that in an RRS sequence, R, if both W(0) and W(1) have parities '0' then every W(n) in R has a parity '0'; and if either or both W(0) and W(1) have parities '1', then one-third of the RRS words in R have parities '0' and two-thirds of the RRS words in R have parities '1', and consequently the period of R is a multiple of 3.

ESTIMATION AND CONTROL PERFORMANCE FOR SPACE-TIME POINT-PROCESS OBSERVATIONS*.

IAN B. RHODES** and DONALD L. SNYDER***

ABSTRACT

Estimation and control problems are examined for a class of models involving a linear system, a quadratic cost, and observations that include a space-time point process as well as the familiar "signal in additive Wiener process" measurements. Motivation for this class of models is given in terms of position sensing and tracking for quantum-limited optical communication problems. These models include as special cases several simpler ones considered previously. As in the simpler cases, the optimum estimator is finite-dimensional and nonlinear, and the optimum controller separates into the optimum estimator followed by the certainty-equivalent control law.

Although the optimum estimator and the optimum controller are finite-dimensional, the corresponding expected error covariance and optimum cost require infinite-dimensional calculations. This motivates the derivation of easily-computed upper and lower bounds on estimator and controller performance. The upper bounds are derived by evaluating exactly the performance of a parametrized family of suboptimum designs; one of these is identified as having smaller performance than any other, thus providing a minimal upper bound within this family. The lower bounds are obtained directly by calculations involving inequalities.

---

*This work was supported in part by the Office of Naval Research under Contract N00014-76-C-0667, in part by the National Science Foundation under Research Grants ENG74-07800 and ENG76-11565, and in part by the National Institutes of Health under Research Grant RR00396 from the Division of Research Resources.

**Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130.

***Department of Electrical Engineering and Biomedical Computer Laboratory, Washington University, St. Louis, Missouri 63130.

I. INTRODUCTION

Snyder and Fishman [1] have considered the problem of estimating the Gaussian state of a linear stochastic system from observations of a point process in which each point has both a spatial and a temporal co-ordinate. The state of the system influences the *spatial* component of the intensity of the observed space-time point process: at any given time, the contours of constant spatial intensity are ellipsoids whose common centroid depends linearly on the current system state. The *temporal* component of the intensity is assumed in [1] to be deterministic. The conditional density of the system state at any time given the past of the observation process is shown to be Gaussian, and the conditional mean and the conditional covariance satisfy finite-dimensional nonlinear stochastic differential equations that are driven by the observed space-time point process.

This model has been generalized in [2], [3] to include causal feedback interactions between the observed point process and the state of the linear stochastic system. Although inclusion of a feedback (control) term destroys the Gaussian-ness of the system state process, it does not alter either the Gaussian form of the *conditional* density of the state given past observations or the finite-dimensionality of the stochastic differential equations for the conditional mean and the conditional covariance. These and related properties underly the derivation of a separation theorem for a stochastic optimal control problem involving these system and observation processes and a quadratic cost functional. Motivation for this stochastic control problem is given in [2], [3] in terms of position sensing and tracking for quantum-limited optical communication problems.

In this paper, we first generalize the model of [2], [3] in two ways. On the one hand, the space-time point process observations are supplemented by continuous observations of a linear function of the system state in an additive Wiener process. The optimum estimator for a restricted version of this problem is included in the dissertation [4] and a corresponding separation theorem is to be included in a forthcoming paper [5]. Here we remove the requirement in [4], [5] that the supplementary observations have the same dimensions as the spatial component of the space-time point process. On the other hand, we allow the *temporal* component of the intensity of the observed space-time point process to be itself a random process. Under appropriate independence assumptions, it is shown that the joint problem of estimating the state of the system and the temporal intensity reduces to two separate problems, one of which is that considered in [2], [3] while the other is a standard estimation problem for point process observations having no spatial component, as discussed, e.g., in [6]. All properties needed to extend the separation theorem for stochastic control problems are retained. These two generalizations are discussed later in terms of the optical position-sensing and tracking problem that motivated [2], [3].

Second, we examine estimation and control performance via upper and lower bounds. While in all cases the optimum estimator and the corresponding conditional error covariance satisfy finite-dimensional stochastic differential equations and thus can be computed on-line, both depend on the observed space-time point process and cannot be precomputed. Insofar as the conditional covariance is concerned, this contrasts with the precomputability that holds for the Kalman filter. One is therefore led to consider the expectation of the conditional covariance, both as a natural measure of estimation performance in its own right and because it happens to be the particular measure of estimation performance that determines the optimum cost in the stochastic control problems considered here and in [2], [3]. However, while the expectation of the conditional covariance is deterministic and in principle can be precalculated, this calculation turns out to be infinite-dimensional. With this in mind, we derive in Sections IV and V easily-precalculable matrix-ordering upper and lower bounds on the expected conditional covariance. The upper bounds are obtained by determining the exact performance of each estimator in a parametrized family of suboptimal estimators whose structure is similar to that of the optimum estimator but for which the mean-square error is precomputable. From within this class, we identify a particular suboptimum estimator whose mean-square error lies at all times below that of any other in the matrix ordering sense. The lower bound is obtained directly using differential and other inequalities.

II. FORMULATION OF THE ESTIMATION AND CONTROL PROBLEMS

Consider the stochastic linear system

$$dx_t = F(t)x_t dt + G(t)u_t dt + V(t)dv_t \tag{1a}$$

$$dz_t = C(t)x_t dt + dw_t; \quad z_0 = 0 \tag{1b}$$

where the state x_t is an n-dimensional random vector, the control u_t is a k-dimensional vector whose measurability is defined later, v and w are independent (normalized) ℓ- and q-dimensional Wiener processes, the random initial state x_0 of (1a) is independent of v and w and is Gaussian with mean $\bar{x}_0$ and covariance Σ_0, and the deterministic uniformly bounded matrix-valued time functions $F(\cdot)$, $G(\cdot)$, $V(\cdot)$ and $C(\cdot)$ have the appropriate dimensions.

In addition to observations of the process z, there are also available observations of a space-time point process defined on $[0,\infty) \times R^m$ as follows. Each point occurrence is identified by a temporal co-ordinate $t \in [0,\infty)$ and a spatial co-ordinate $r \in R^m$. Let τ and A be Borel sets in $[0,\infty)$ and R^m, respectively, and denote by $N(\tau \times A)$ the number of points occurring in $\tau \times A$. We define $N_t = N([0,t) \times R^m)$ to be the number of points up to but not including time t regardless of their spatial location; $N_\cdot$ is taken to be a Poisson counting process with intensity μ, where μ is a stochastic process that is independent of x_0, v and w, and μ_t is almost-surely positive. Given that N has a jump at t (i.e. $N_{t-} \neq N_{t+}$), the spatial location r of the point is taken to be an m-dimensional Gaussian random vector with mean $H(t)x_t$ and known positive definite covariance $R(t)$, where $H(\cdot)$ is a known $m \times n$-matrix valued time function. Given N_s and x_s for $s \geq 0$, the spatial locations are independent random vectors that are independent of all other random entities. Thus the space-time point process can be thought of as having an intensity

$$\lambda_t(r,x_t,\mu_t) = \mu_t \gamma_t(r,x_t) \tag{2}$$

that separates into the product of a temporal component μ_t that underlies the Poisson counting process N and a spatial component

$$\gamma_t(r,x_t) \sim N(H(t)x_t,R(t)) = (2\pi)^{-m/2}[\det R(t)]^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(r-H(t)x_t)' R^{-1}(t)(r-H(t)x_t)\}$$

that gives the density of the spatial location r of the point occurring at t.

Let $(\Omega,\mathcal{F},P)$ be the underlying probability space. We denote by $\mathcal{Z}_t$ the the sub-σ-algebra of $\mathcal{F}$ generated by the process z over the interval $[0,t)$, and by $\mathcal{N}_t$ the sub-σ-algebra generated by the space-time point process over $[0,t)$. Let $\mathcal{B}_t = \mathcal{Z}_t \vee \mathcal{N}_t$, the σ-algebra generated by $\mathcal{Z}_t$ and $\mathcal{N}_t$. It is assumed throughout that u_t is $\mathcal{B}_t$-measurable and such that the solution to (1a) is well-defined; such controls will be henceforth called *admissible*.

The estimation problem to which we address ourselves is to find the conditional means

$$\hat{x}_t \triangleq E[x_t|\mathcal{B}_t], \quad \hat{\mu}_t \triangleq E[\mu_t|\mathcal{B}_t] \tag{3}$$

and the corresponding conditional covariances

$$\Sigma_t \triangleq \mathrm{cov}[x_t|\mathcal{B}_t], \quad \Gamma_t \triangleq \mathrm{cov}[\mu_t|\mathcal{B}_t]. \tag{4}$$

The control problem we examine is to find the admissible control $\{u_t: t \in [0,T]\}$ that minimizes the quadratic cost functional

$$J[u] = E\{\int_0^T [u_t'P(t)u_t + x_t'Q(t)x_t]dt + x_T'Sx_T\} \tag{5}$$

where the symmetric uniformly-bounded matrix-valued time functions have the appropriate dimensions with $Q(t)$ and S non-negative definite and $P(t)$ positive difinite.

Our notation is generally as follows: lower case letters denote vectors, upper case letters denote matrices, and script letters denote σ-algebras; v_t denotes a time-indexed random vector, in contrast to $v(t)$ which denotes a time-indexed deterministic vector; everything takes place on the fixed, finite time interval $[0,T]$; $y \sim N(q,Q)$ means that y is Gaussian with mean q and covariance Q; the inequality $P \leq Q$ between symmetric, non-negative definite matrices means that $Q-P$ is non-negative definite.

III. SOLUTION OF THE OPTIMAL ESTIMATION AND CONTROL PROBLEMS

Theorem 1. The conditional density of x_t given $\mathcal{B}_t$ is Gaussian and the conditional mean and the conditional covariance satisfy the finite-dimensional nonlinear stochastic differential equations:

$$\begin{aligned} d\hat{x}_t &= F(t)\hat{x}_t dt + G(t)u_t dt + \Sigma_t C'(t)[dz_t - C(t)\hat{x}_t dt] \\ &\quad + \int_{R^m} M_t[r - H(t)\hat{x}_t]N(dt \times dr); \quad \hat{x}_0 = E[x_0] \end{aligned} \tag{6}$$

$$\begin{aligned} d\Sigma_t &= F(t)\Sigma_t dt + \Sigma_t F'(t)dt + V(t)V'(t)dt - \Sigma_t C'(t)C(t)\Sigma_t dt \\ &\quad - M_t H(t)\Sigma_t dN_t; \quad \Sigma_0 = \mathrm{cov}[x_0] \end{aligned} \tag{7}$$

where

$$M_t = \Sigma_t H'(t)[H(t)\Sigma_t H'(t) + R(t)]^{-1} \tag{8}$$

If $\mathrm{cov}[x_0]$ is positive definite then Σ_t is almost-surely positive definite and its inverse satisfies the finite-dimensional nonlinear stochastic differential equation

$$\begin{aligned} d\Sigma_t^{-1} &= -\Sigma_t^{-1}F(t)dt - F'(t)\Sigma_t^{-1}dt - \Sigma_t^{-1}V(t)V'(t)\Sigma_t^{-1}dt + C'(t)C(t)dt \\ &\quad + H'(t)R^{-1}(t)H(t)dN_t; \quad \Sigma_0^{-1} = (\mathrm{cov}[x_0])^{-1} \end{aligned} \tag{9}$$

The conditional density of μ_t given $\mathcal{B}_t$ coincides with the conditional density of μ_t given the σ-algebra $\mathcal{T}_t$ generated by the past of the process N_t, i.e. $f_t(\mu|\mathcal{B}_t) = f_t(\mu|\mathcal{T}_t)$, assuming the control u_t is satisfies a technical property specified in the proof and discussed immediately thereafter.

Proof. The derivation of (6) - (9) we give here parallels the proofs of Lemma 1 and Proposition 1 in [1] which establish the corresponding result for the special case where $u_t \equiv 0$, $\mu_t = \mu(t)$ is deterministic, and $C(t) \equiv 0$ (i.e. the observations z are not available). In outline, the modifications that are made to include each of these generalizations are as follows: the introduction of a $\mathcal{B}_t$-measurable u_t causes no difficulty since u_t is deterministic in all calculations which involve probability measures conditioned on $\mathcal{B}_t$; the presence of nonzero $C(\cdot)$ merely adds an additional term that is familiar for this "signal in Wiener process" observation model; the generalization to random μ_t is handled by temporarily conditioning everything on the σ-algebra $\mathcal{M}$ generated by μ over [0,T] and subsequently finding that the stochastic differential equation for the conditional density of x_t given $\mathcal{B}_t$ and $\mathcal{M}$ turns out to be independent of $\mathcal{M}$.

Letting $\phi_t = \exp[jy'x_t]$, where $y \in R^n$ is nonrandom, we find using the Itô rule that

$$d\phi_t = \phi_t \psi_t dt + \phi jy'V(t)dv_t$$

where

$$\psi_t = jy'[F(t)x_t + G(t)u_t] - \tfrac{1}{2}y'V(t)V'(t)y$$

Letting $\mathcal{F}_t = \mathcal{B}_t \vee \mathcal{M}$ and defining for the moment $\hat{x}_t = E[x_t | \mathcal{F}_t]$ and $\hat{\lambda}_t(r) = E[\lambda_t(r,x_t,\mu_t) | \mathcal{F}_t]$, it follows from our standing assumptions that, for any Borel set $B \in R^m$, $dz_t - C(t)\hat{x}_t dt$ and $N(dt \times B) - \int_B \hat{\lambda}_t(r)dr\,dt$ are independent, independent-increment processes relative to $\mathcal{F}_t$. We then have, analogously to [1, Eq. 9], that the conditional characteristic function $\hat{M}_t(jy) = E[\phi_t | \mathcal{B}_t \vee \mathcal{M}]$ of x_t given $\mathcal{B}_t \vee \mathcal{M}$ satisfies

$$\begin{aligned} d\hat{M}_t(jy) = {} & E\{\phi_t \psi_t | \mathcal{F}_t\}dt + E\{\phi_t (x_t - \hat{x}_t)' | \mathcal{F}_t\} C'(t)[dz_t - C(t)\hat{x}_t dt] \\ & + \int_{R^m} E\{\phi_t [\lambda_t(r,x_t,\mu_t) - \hat{\lambda}_t(r)] | \mathcal{F}_t\} \\ & \times \hat{\lambda}_t^{-1}(r)[N(dt \times dr) - \hat{\lambda}_t(r)dr\,dt] \end{aligned}$$

Taking inverse Fourier transforms and simplifying then yields the following stochastic differential equation for the conditional density of x_t given $\mathcal{F}_t$ (c.f. [1, Eq. 5])

$$\begin{aligned} dp_t(X|\mathcal{B}_t \vee \mathcal{M}) = {} & L[p_t(X|\mathcal{F}_t)]dt + p_t(X|\mathcal{F}_t)[X - \hat{x}_t]'C'(t)[dz_t - C(t)\hat{x}_t dt] \\ & + p_t(X|\mathcal{F}_t) \int_{R^m} [\lambda_t(r,X,\mu_t) - \hat{\lambda}_t(r)]\hat{\lambda}_t^{-1}(r)N(dt \times dr) \quad (10) \end{aligned}$$

where

$$\begin{aligned} L[q] = {} & - \sum_{i=1}^{n} \partial[(F(t)X + G(t)u_t)q]_i / \partial X_i \\ & + \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \partial^2 [V(t)V(t)q]_{ij} / \partial X_i \partial X_j \end{aligned}$$

Recalling from (2) that $\lambda_t(r,x_t,\mu_t) = \mu_t \cdot \gamma_t(r,x_t)$, we see that the integrand of the last term in (10) can be rewritten as $[\gamma_t(r,X) - \hat{\gamma}_t(r)]\hat{\gamma}_t^{-1}(r)$, where $\hat{\gamma}_t(r) = E[\gamma_t(r,x_t)|\mathcal{F}_t]$. Noting that $\hat{\gamma}_t(r)$ and $\hat{x}_t$ can be written as integrals involving $p_t(X|\mathcal{F}_t)$, the evolution in time of (10) does not depend in in any way on μ_t. Thus $p_t(X|\mathcal{B}_t \vee \mathcal{M}) = p_t(X|\mathcal{B}_t)$, and (10) can be rewritten as

$$dp_t(X|\mathcal{B}_t) = \mathcal{L}[p_t(X|\mathcal{B}_t)]dt + p_t(X|\mathcal{B}_t)[X - \hat{x}_t]C'(t)[dz_t - C(t)\hat{x}_t dt]$$

$$+ p_t(X|\mathcal{B}_t) \int_{R^m}[\gamma_t(r,X) - \hat{\gamma}_t(r)]\hat{\gamma}_t^{-1}(r)N(dt \times dr) \tag{11}$$

with the Gaussian distribution $N(\bar{x}_0,\bar{\Sigma}_0)$ as initial condition. Of course, $E[x_t|\mathcal{B}_t \vee \mathcal{M}] = E[x_t|\mathcal{B}_t]$, so the definition of x_t given in the statement of Theorem 1 coincides with the temporary definition introduced in the proof; similar remarks apply to $\hat{\gamma}_t$.

The proof that $p_t(X|\mathcal{B}_t)$ is Gaussian with mean $\hat{x}_t$ given by (6) and covariance Σ_t given by (7) or (9) then follows from a straightforward inductive proof similar to that of Proposition 1 in [1]: in the intervals between point occurrences of the space-time point process, $p_t(X|\mathcal{B}_t)$ evolves according to the first two terms on the right side of (11); this is simply Kushner's equation for linear system (1a) with linear observations (1b), and is known to yield a conditional density $p_t(X|\mathcal{B}_t)$ that is Gaussian with mean $\hat{x}_t$ and covariance Σ_t satisfying (6) and (7) or (9), respectively, with the last term on the right side of each deleted. At those instants when a point occurs in the space-time point process, a jump occurs in $p_t(X|\mathcal{B}_t)$ because of the last term on the right side of (11). However, it turns out that $p_t(X|\mathcal{B}_t)$ remains Gaussian after this jump because it was Gaussian before the jump and because the spatial intensity $\gamma_t(r,X)$ is Gaussian. As in [1], calculation of the last term on the right side of (11) shows that the jump in the conditional mean is given by the last term on the right side of (6), the jump in conditional covariance by the last term on the right side of (7), and the jump in the inverse of the conditional covariance by the last term on the right side of (9).

Finally, to prove that $f_t[\mu|\mathcal{B}_t] = f[\mu|\mathcal{T}_t]$, let x^* satisfy $dx_t^* = F(t)x_t^*dt + G(t)u_t dt$; $x_0^* = E[x_0]$. Then $\tilde{x}_t \triangleq x_t - x_t^*$ and $\tilde{z}_t \triangleq z_t - C(t)x_t^*$ satisfy

$$d\tilde{x}_t = F(t)\tilde{x}_t dt + V(t)dv_t; \quad d\tilde{z}_t = C(t)\tilde{x}_t dt + dw_t \tag{12a}$$

with $\tilde{x}_0 \sim N(0,\bar{\Sigma}_0)$ and $z_0=0$, while $\tilde{r} \triangleq r-Hx^*$ has spatial intensity

$$\tilde{\gamma}_t(\tilde{r},\tilde{x}_t) \sim N(H(t)\tilde{x}_t,R(t)). \tag{12b}$$

Let $\tilde{\mathcal{Z}}_t$ be the σ-algebra generated by $\tilde{z}$ over $[0,t)$ and let $\tilde{\mathcal{N}}_t$ be that generated by the space-time process that is obtained from the original one by leaving N. unchanged and replacing r by $r-Hx^*$. Then, *under the assumption that* μ_t *is also* $\mathcal{Z}_t \vee \tilde{\mathcal{N}}_t$*-measurable*, an argument that parallels the proof of Lemma 1 in [7] shows that $\mathcal{Z}_t \vee \mathcal{N}_t = \tilde{\mathcal{Z}}_t \vee \tilde{\mathcal{N}}_t$. This assumption is discussed shortly. Thus, it is equivalent to prove that $f_t[\mu|\tilde{\mathcal{Z}}_t \vee \tilde{\mathcal{N}}_t] = f_t[\mu|\mathcal{T}_t]$. Now, because μ and N are independent of x_0, v and w, so also are they independent of $\tilde{x}$ and $\tilde{z}$; thus, the joint density of μ and the event that $t_1,t_2,\ldots,t_k$ are the occurrence times of N over $[0,t)$ satisfies $f[\mu,t_1,\ldots,t_k|\tilde{\mathcal{Z}}_t] = f[\mu,t_1,..,t_k]$. Equivalently,

$$f[\mu|\mathcal{T}_t \vee \tilde{\mathcal{Z}}_t]f[t_1,..,t_k|\tilde{\mathcal{Z}}_t] = f[\mu|\mathcal{T}_t]f[t_1,..,t_k]$$

Because N and $\tilde{z}$ are independent, $f[t_1,..,t_k|\tilde{\mathcal{Z}}_t] = f[t_1,..t_k]$ and therefore

$$f[\mu|\mathcal{T}_t \vee \tilde{\mathcal{Z}}_t] = f[\mu|\mathcal{T}_t] \tag{12c}$$

Also because the spatial components $\tilde{r}_1,\tilde{r}_2,\ldots\tilde{r}_{N_t}$ are independent of μ given $\tilde{\mathcal{X}}_t$ and $\mathcal{T}_t$, we have

$$f[\mu,\tilde{r}_1,\ldots,\tilde{r}_{N_t}|\mathcal{T}_t \vee \tilde{\mathcal{X}}_t] = f[\mu|\mathcal{T}_t \vee \tilde{\mathcal{X}}_t]f[\tilde{r}_1,\ldots,\tilde{r}_{N_t}|\mathcal{T}_t \vee \tilde{\mathcal{X}}_t]$$

where $\tilde{\mathcal{X}}_t$ is the σ-algebra generated by $\tilde{x}$ over $[0,t)$. Replacing $\tilde{\mathcal{Z}}_t$ by $\tilde{\mathcal{X}}_t$ in the argument leading to (12c), we have $f[\mu|\mathcal{T}_t \vee \mathcal{X}_t] = f[\mu|\mathcal{T}_t]$ for the first term on the right side, while the left side can be written $f[\mu|\tilde{r}_1,\dots,\tilde{r}_{N_t}, \mathcal{T}_t \vee \mathcal{X}_t]f(r_1,\dots,r_{N_t}|\mathcal{T}_t \vee \tilde{\mathcal{X}}_t]$. Cancelling the second term on each side, we are left with

$$(f[\mu|\tilde{\mathcal{N}}_t \vee \tilde{\mathcal{X}}_t]=)f[\mu|\tilde{r}_1,\dots,\tilde{r}_{N_t},\ \mathcal{T}_t \vee \tilde{\mathcal{X}}_t] = f[\mu|\mathcal{T}_t],$$

which in combination with (12c) gives the required result.

Remark 1. The technical assumption that u_t is $\tilde{\mathcal{N}}_t \vee \tilde{\mathcal{Z}}_t$-measurable which is required for the proof that $f_t[\mu|\mathcal{B}_t] = f_t[\mu|\mathcal{T}_t]$ also arises as a sufficient condition in [7]. A generalization of [7, Theorem 3] shows that this will be the case if u is generated from the past of z, N and r or $\tilde{z}$, N and $\tilde{r}$ using a suitably smooth control law. Specifically, it will be so if μ is generated as a Lipschitz function of the state of a suitably smooth finite-dimensional system; included here, in particular, is a control u_t so generated from $\hat{x}_t$ of (6), which is of interest because this is the case for the optimum control found later.

Remark 2. The stochastic differential equations for $\hat{x}_t$ and Σ_t given in Theorem 1 admit an intuitively simple interpretation. In the intervals between point occurrences of the space-time process, the problem reduces to one of estimating x_t from the observations z; the non-occurrence of further points in these intervals provides information about μ_t but none about x_t because of the separability (2) of λ and our standing assumptions concerning independence. Thus, during these intervals we are left with the standard Kalman filtering problem of estimating the state of the linear system (1a) from the observations (1b); if x_t is conditionally Gaussian at the beginning of each such interval, it remains so throughout with mean and covariance which evolve according to (6) and (8) or (9) with the last term deleted in each. (This, or course, is also reflected in the equation (11) for the conditional density reducing to Kushner's equation during these intervals.) We now observe that x_t is, in fact, conditionally Gaussian at the beginning of each such interval because it is at t=0 and because it remains Gaussian after each point occurrence: indeed, at each occurrence (t,r) of the space-time point process the spatial observation r is an independent observation on a Gaussian random variable with mean Hx_t and covariance R; this is equivalent to a discrete observation of the form

$$r = H(t)x_t + \xi$$

where $\xi \sim N(0,R)$ is independent of x and z. Thus, from standard estimation theory for Gaussian random variables [e.g. 10], the conditional density remains Gaussian and the change in conditional mean and covariance of x_t after accounting for this new observation are

$$d\hat{x}_t \triangleq \hat{x}_{t+} - \hat{x}_t = \Sigma_t H'[H\Sigma H' + R]^{-1}(r - H\hat{x}_t) = M_t(r - H\hat{x}_t)$$

$$d\Sigma_t \triangleq \Sigma_{t+} - \Sigma_t = -\Sigma_t H'[H\Sigma H' + R]^{-1}H\Sigma_t = -M_t H\Sigma_t$$

$$d\Sigma_t^{-1} \triangleq \Sigma_{t+}^{-1} - \Sigma_t^{-1} = H'R^{-1}H$$

Of course, this term is to be included only when an occurrence takes place at (t,r); multiplying each of these expressions by N(dt × dr) and integrating over R^m takes care of this, and constitues the last term in (6), (7) and (9), respectively.

Remark 3. It was observed in the proof that $p_t(x|\mathcal{B}_t \vee \mathcal{M})$ does not depend on $\mathcal{M}$, and thus x_t and μ are conditionally independent given $\mathcal{B}_t$; in particular, x_t and μ_t are conditionally independent given $\mathcal{B}_t$, and thus the joint problem of estimating x_t and μ_t given $\mathcal{B}_t$ separates into two separate problems of estimating x_t and estimating μ_t:

$$f(x_t,\mu_t|\mathcal{B}_t) = f(x_t|\mathcal{B}_t)f(\mu_t|\mathcal{B}_t)$$

Furthermore, the final part of Theorem 1 establishes that $f(\mu_t|\mathcal{B}_t)=f(\mu_t|\mathcal{T}_t)$ depends only on the Poisson counting process N and does not depend on z or the spatial locations of the points. This latter estimation problem with various models for μ is examined, for example, in [6].

Theorem 2. The unique admissible control u_t° that minimizes the cost (5) is given by

$$u_t^\circ = -P^{-1}(t)G'(t)K(t)\hat{x}_t \triangleq -L(t)\hat{x}_t \tag{13}$$

where $\hat{x}_t = E[x_t|\mathcal{B}_t]$ satisfies the finite-dimensional nonlinear stochastic differential equation (6) with Σ_t given by (7), and the n × n symmetric non-negative definite matrix K(t) satisfies the Riccati equation

$$\dot{K}(t) = -K(t)F(t) - F'(t)K(t) + K(t)G(t)P^{-1}(t)G'(t)K(t) - Q(t); \; K(T)=S \tag{14}$$

The corresponding minimum value of J is

$$J[u^\circ] = E\{x_0'K(0)x_0\} + \int_0^T tr[KGP^{-1}G'K\, E\{\Sigma_t\} + KVV']dt, \tag{15}$$

where tr denotes trace.

Proof. According to Åström [8], J[u] can be rewritten as

$$J[u] = \int_0^T E\{||u_t + L(t)x_t||^2_{P(t)}\}dt + [\text{Right side of (15)}]$$

where $\hat{x}_t = E[x_t|\mathcal{B}_t]$ and $||y||_P^2 = y'Py$. The first term on the right side is non-negative, and zero if and only if $u_t = -L(t)\hat{x}_t$. Thus (13) gives the unique optimum control provided the right side of (15) is invariant under changes in u. The only way for (15) to be u-dependent is through $E\{\Sigma_t\}$, and (7) shows that the only possibility for Σ_t to vary with u is via N_t. But N_t is a Poisson counting process with rate μ_t, and both N_t and μ_t are specified at the outset as mappings on $(\Omega,\mathcal{F},P)$ without any reference to u. Hence Σ_t and, therefore, $E\{\Sigma_t\}$ and the right side of [15] are invariant under changes in u, and the proof is complete.

Remark 4. Theorem 2 shows that the solution to this stochastic control problem can be realized with a spearated estimator-controller in which the estimator is nonlinear, mean-square optimal, and finite-dimensional and the controller is the certainty-equivalent linear control law (i.e. the optimum linear control law for the deterministic problem in which $x_0=E[x_0]$, $v_t = E[v_t]\equiv 0$, and x_t is known exactly). This result, therefore, includes as special eases the familiar linear-quadratic-Gaussian "Separation Theorem" (where the space-time point process observations are absent) [e.g. 8] and the similar results in [1] - [5] for restricted versions of the space-time point process observations.

We observe that Σ_t, the conditional covariance of x_t given $\mathcal{B}_t$, is not precomputable because the last term on the right side of (7) depends on the particular realization of the counting process N_t. One is, therefore, led to consider $\overline{\Sigma}(t) \triangleq E[\Sigma_t]$, both as a natural measure of estimation performance in its own right and as the particular measure of estimation performance that determines the optimum control cost (15). However, while $\overline{\Sigma}(t)$

is deterministic and in principle can be precalculated, this calculation is infinite-dimensional. One way of seeing this is to observe that an attempt to calculate $\bar{\Sigma}(t)$ by taking expectations of both sides of (7) is complicated by the last term on the right side, which requires the calculation of $E\{\Sigma H'[H\Sigma H' + R]^{-1}H\Sigma\}$. While a differential equation for this can be written down, it, in turn, requires expectations of additional nonlinear functions of Σ_t, and so on *ad infinitum* in a mushrooming requirement for additional terms that is familiar from other nonlinear filtering situations. Accordingly, we turn our attention to deriving easily-computed upper and lower bounds on $\bar{\Sigma}(t)$. These estimation bounds then directly imply upper and lower bounds on the optimum control performance (15).

IV. SUBOPTIMUM ESTIMATORS AND UPPER BOUNDS

Our approach to finding easily-computed upper bounds on $E[\Sigma_t]$ is to examine a parametrized family of suboptimum estimators whose mean-square performance can easily be calculated exactly. For each suboptimum estimator the corresponding mean square-error is then trivially a matrix-ordering upper bound on $E[\Sigma_t]$. Furthermore, we show that there exists a suboptimum estimator in this family whose mean-square performance is at all times smaller than that of any other, thus providing a minimal upper bound within this class.

Motivated by the form of the optimum estimator (6), we consider the family of suboptimum estimators

$$dx_t^{\#} = F(t)x_t^{\#}dt + G(t)u_t dt + N(t)[dz_t - C(t)x_t^{\#}dt]$$
$$+\int_{R^m} M(t)[r - H(t)x_t^{\#}]N(dt \times dr) \qquad (16)$$

parametrized by the deterministic uniformly-bounded n × m- and n × q-matrix valued time functions $N(\cdot)$ and $M(\cdot)$. This family does not include the optimum estimator (6) in which M_t is a random matrix which depends on N through Σ. Apart from the requirement that $M(\cdot)$, $N(\cdot)$ be deterministic, the sub-optimum estimator (16) and the optimum estimator (6) share the same structure. The nonrandomness of M enables us to write down an ordinary n × m-matrix differential equation for the mean square error of the sub-optimum estimator (16). Indeed, subtracting (16) from (1a), it follows directly by straightforward calculation that

$$S(t) \triangleq E[(x_t - x_t^{\#})(x_t - x_t^{\#})'] \qquad (17)$$

satisfies the linear matrix differential equation

$$\dot{S} = (F - NC)S + S(F - NC)' + VV' + NN'$$
$$+ \bar{\mu}\{M[HSH' + R]M - MHS - SH'M'\};\ S(0) = \text{cov}[x_0] \qquad (18)$$

where we have suppressed the common argument t of all entries, and where $\bar{\mu}(t) = E[\mu_t]$. Because all coefficients in (18) are uniformly bounded, a unique solution to (18) exists for all $t \in [0,\infty)$. We thus have proved:
Theorem 3. For any uniformly bounded $M(\cdot)$ and $N(\cdot)$, the mean-square performance (17) of the estimator (16) satisfies the linear matrix differential equation (18), and this is a matrix ordering upper bound on $E[\Sigma_t]$, i.e. for all $t \in [0,\infty)$,

$$\bar{\Sigma}(t) = E[\Sigma_t] \leq S(t), \qquad (19)$$

in the sense that $S(t) - \bar{\Sigma}(t)$ is non-negative definite.

We now show that there exists a choice of $M(\cdot)$ and $N(\cdot)$ in (16) for which the corresponding mean-square performance given by (18) lies at all times below that for any other choice.

Theorem 4. Let $S^*(\cdot)$, $M^*(\cdot)$ and $N^*(\cdot)$ satisfy

$$S^* = FS^* + S^*F' + VV' - S^*C'CS^* - \bar{\mu}S^*H'[HS^*H' + R]^{-1}HS^*;\ S^*(0)=\mathrm{cov}[x_0] \tag{20}$$

$$M^* = S^*H'[HS^*H' + R]^{-1},\quad N^* = S^*C' \tag{21}$$

where the common argument t of all entries in (20) and (21) is suppressed. Let $S(\cdot)$ be the solution to (18) for some arbitrary $M(\cdot)$ and $N(\cdot)$. Then, for all $t \in [0,\infty)$,

$$E[\Sigma_t] \leq S^*(t) \leq S(t) \tag{22}$$

and $S^*(t) = E\{(x_t - x_t^*)(x_t - x_t^*)'\}$ is the mean-square performance of the bound-minimal estimator

$$dx_t^* = Fx_t^*dt + Gu_tdt + S^*C'[dz_t - Cx_t^*dt] + S^*H'[HS^*H' + R]^{-1}\int_{R^m}[r - Hx_t^*]N(dt \times dr) \tag{23}$$

Proof. Completing the square on the right side of (18) yields

$$\dot{S} = FS + SF' + VV' - SC'CS - \bar{\mu}SH'[HSH' + R]^{-1}HS + (N - SC')(N - SC')' + \bar{\mu}[M - SH'(HSH' + R)^{-1}](HSH' + R)[M - SH'(HSH' + R)^{-1}]',\quad S(0) = \mathrm{cov}[x_0] \tag{24}$$

For given $S(t)$, the right side of (23) is clearly minimized by making the last two non-negative definite terms 0, in which case the right side of (23) reduces to that of (20) while the minimizing choices of $M(t)$ and $N(t)$ are given by (21). It remains to show that this instantaneous ordering on the time derivative produces a permanent ordering of the solutions over $[0,T]$, i.e. that the solution to (20) lies at all times below that of (23) in the matrix ordering. This is readily accomplished by using Lemma 1 in (9), after appropriate modifications to reflect that initial conditions, rather than final conditions, are of interest here. This means that in [9, Lemma 1] the left sides of (*) and (**) should be replaced by $+\dot{X}$ and $+\dot{Y}$, respectively, and all time orderings $0 \leq t \leq s \leq T$ replaced by $0 \leq s \leq t \leq 1$. Then, letting (24) play the role of (*) and (20) the role of (**) and checking conditions 1) to 4) of [9, Lemma 1] we have: 1) and 2) are trivial under our standing assumptions; 3) holds because, by a subsidiary application of [9, Lemma], the solution to (20) lies at all times above that of the Riccati equation

$$\dot{\Gamma} = F\Gamma + \Gamma F' + VV' - \Gamma[C'C + \bar{\mu}H'R^{-1}H]\ ;\quad \Gamma(0) = \mathrm{cov}[x_0] \tag{25}$$

for which 3) is known to hold; finally, 4) holds for (18) and therefore (24) because if $S_1(t)$ and $S_2(t)$ are the respective solutions to (18) with initial conditions $S_1(0)$ and $S_2(0)$, $S_1(0) \geq S_2(0)$, then $S_1(t) \geq S_2(t)$ for all t, since

$$\dot{S}_1 - \dot{S}_2 = (F - NC - \bar{\mu}MH)(S_1 - S_2) + (S_1 - S_2)(F - NC - \bar{\mu}MH)' + \bar{\mu}MH(S_1 - S_2)H'M;\quad (S_1 - S_2)(0) = 0$$

and the solution to this lies at all times above the (identically zero) solution to

$$\dot{Y} = (F - NC - \bar{\mu}MH)Y + Y(F - NC - \bar{\mu}MH)'$$

by a further subsidiary application of [9, Lemma 1].

Remark 5. The evaluation of the performance of the suboptimum estimator (16) is a second-order analysis that uses only the means and covariances of the various random variables and processes and makes no use of the Gaussian-ness of x_0, v, w and r. Thus the results of this section remain valid if v and w are replaced by normalized uncorrelated-increment processes that are uncorrelated with each other and with x_0 and r, with x_0 having any distribution with mean $\bar{x}_0$ and covariance Σ_0 and the spatial intensity $\gamma_t(r,x_t)$ being any distribution with mean Hx and covariance R such that r is uncorrelated with x_0. The bound-minimal estimator (23) can then be viewed as the best estimator in the family (16). These estimators are bilinear because of the product r.N(dt × dr) in the last term, though they might also be considered to be in a sense linear, to the extent that N(dt × dr) merely signals the arrival of a spatial observation r which, as with z, is utilized linearly in the production of $\hat{x}$.

V. ESTIMATION LOWER BOUNDS AND CONTROL BOUNDS

Theorem 5. Let S_* be the solution to (25). Then, for all $t \geq 0$, $S_*(t)$ is a matrix-ordering lower bound on $E[\Sigma_t]$, i.e.

$$S_*(t) \leq E[\Sigma_t] \tag{26}$$

The corresponding lower bound on the optimum control cost is

$$J[u^\circ] \geq E\{x_0'K(0)x_0\} + \int_0^T \mathrm{tr}[KGP^{-1}G'KS_* + KVV']dt \tag{27}$$

Proof. We have from (9) than $\overline{\Sigma^{-1}} \triangleq E[\Sigma_t^{-1}]$ satisfies

$$\dot{\overline{\Sigma^{-1}}} = -\overline{\Sigma^{-1}}\,F - F'\,\overline{\Sigma^{-1}} - E[\Sigma^{-1}VV'\Sigma^{-1}] + \bar{\mu}H'R^{-1}H; \quad \overline{\Sigma^{-1}}(0) = (\mathrm{cov}[x_0])^{-1}$$

$$= -\overline{\Sigma^{-1}}\,F - F'\,\overline{\Sigma^{-1}} - \overline{\Sigma^{-1}}\,VV'\,\overline{\Sigma^{-1}} + \bar{\mu}H'R^{-1}H - \Delta \tag{28}$$

where

$$\Delta = E[\Sigma^{-1}VV'\Sigma^{-1}] - \overline{\Sigma^{-1}}\,VV'\,\overline{\Sigma^{-1}} = \mathrm{cov}[\Sigma^{-1}V] \geq 0 \tag{29}$$

It then follows from [9, Lemma 1] that $\overline{\Sigma^{-1}}$ lies at all times below the solution to

$$\dot{\Xi} = -\Xi F - F'\Xi - \Xi VV'\Xi + \bar{\mu}H'R^{-1}H; \quad \Xi(0) = (\mathrm{cov}[x_0])^{-1} \tag{30}$$

Thus

$$\Xi(t) \geq \overline{\Sigma^{-1}}(t) \triangleq E[\Sigma_t^{-1}] \geq (E\Sigma_t)^{-1}, \tag{31}$$

the last inequality being a matrix version of Jensen's inequality proved in the Appendix. Taking inverses of (31) and noting that if S_* is the solution to (25) then S_*^{-1} is the solution to (30), we have the desired result (26). The control bound (27) then follows by combining (15) and (26).

We remark in passing that (25) gives the covariance of the optimum estimator when the space time point process observations are replaced by continuous observations of the form

$$dy_t = Hx_t dt + (\bar{\mu}^{-1}R)^{1/2}dn_t$$

where n_t is a Wiener process independent of x_0, v and w.

Remark 6. Comparing (20) for the minimal upper bound with (25) for the lower bound, we see that these two bounds will be close to each other and thus to $E[\Sigma_t]$ if HS^*H' is small compared with R (or, equivalently, if $H'R^{-1}H$ is small compared with S^*). Both bounds will also be close to each other and to the optimum performance if the mean intensity μ is small. These are discussed later in terms of our motivating example.

Once we have deduced upper and lower bounds on the estimation performance $E[\Sigma_t]$, corresponding bounds on the optimum control performance follow directly by substitution of these bounds for $E[\Sigma_t]$ in (15):

Theorem 6. Upper and lower bounds on the optimum control performance $J[u°]$ of (15) are

$$E\{x_0'K(0)x_0\} + \int_0^T \mathrm{tr}[KGP^{-1}G'PS_* + KVV']dt$$

$$\leq J[u°] \leq E\{x_0'K(0)x_0\} + \int_0^T \mathrm{tr}[KGP^{-1}G'KS^* + KVV']dt$$

where S_* is the solution to (25) and S^* is the solution to (20).

VI. DISCUSSION

The above estimator-controller solution extends results in [2] and [3] to include a more general form of observation. Just as with the observation model in [2] and [3], this more general observation is motivated by communication systems that employ a narrow beam of light as a carrier, by star tracking systems, and by infra-red tracking systems, all of which have a requirement for position sensing and active tracking to maintain optical alignment in the presence of a variety of disturbances. We shall indicate how the models of [2, Sec. 4] and [3, Sec. 4] are usefully extended by this more general observation. The estimator-controller solution of Theorem 2 provides a possible tool for the design of an optical tracking system under the conditions indicated below, and the performance bounds of Sections IV and V provide the means for predicting the performance of such designs.

Let $I(t,\vec{r})$ denote the light intensity at time $t \in [0,\infty)$ and position $\vec{r} \in \mathcal{R}$ of an optical field incident on the photoemissive surface of a two-dimensional photodetector on boresight and without any motions. Here, $\mathcal{R}$ is a subregion of R^2 corresponding to the photoemissive surface. We assume a Gaussian intensity-profile

$$I(t,\vec{r}) = I_0(t)\exp\{-\tfrac{1}{2}\vec{r}'R^{-1}(t)\vec{r}\}.$$

Vibration, beam steering due to propagation of the light beam through atmospheric turbulence, and other effects cause the spot of light on the photoemissive surface to move about in a random fashion and to fluctuate randomly in optical intensity. In this case, the intensity profile becomes

$$I(t,\vec{r},y_m(t)) = I_0(t)\exp\{-\tfrac{1}{2}[\vec{r}-y_m(t)]'R^{-1}(t)[\vec{r}-y_m(t)]\},$$

where $y_m(t)$ models the random motions, and $I_0(t)$ is a random process (e.g., a lognormal process) that models random intensity fluctuations. We assume that $\{y_m(t);t\geq 0\}$ is derived from a Gaussian diffusion satisfying

$$dx_m(t) = F_m(t)x_m(t)dt + V_m(t)dv_m(t), \quad y_m(t) = H_m(t)x_m(t),$$

where $\{v_m(t);t\geq 0\}$ is a standard Wiener process. The fading process $\{I_0(t); t\geq 0\}$ is assumed to be independent of motion processes but is otherwise arbitrary. The purpose of the tracking controller is to compensate for these random motions and random fading in order to maintain optical alignment. Thus, in the presence of a controller to position telescopes, mirrors, or other pointing devices, the intensity becomes

$$I(t,\vec{r},y_m(t),y_p(t))$$

$$= I_0(t)\exp\{-\tfrac{1}{2}[\vec{r}-y_m(t) + y_p(t)]'R^{-1}(t)[\vec{r}-y_m(t) + y_p(t)]\}$$

where $y_p(t) - y_m(t)$ is the tracking error. Ideally, this error should be zero, but this cannot be accomplished for two reasons: the position error $y_m(t)$ is unknown and must be estimated from data available at the photodetector output, and the tracking devices will have some inertia so that $y_m(t)$ cannot be tracked instantaneously even if it were known. We model the tracking devices by a linear stochastic plant

$$dx_p(t) = F_p(t)x_p(t)dt + G_p(t)u(t)dt + V_p(t)dv_p(t)$$

$$y_p(t) = H_p(t)x_p(t),$$

where u(t) is the input to the tracking devices from the tracking controller, and $\{v_p(t);t\geq 0\}$ is a standard Wiener process modeling local disturbances such as those due to vibration.

Photoelectron conversions take place in the photoemissive surface at a rate proportional to the incident light intenstiy [3]. Thus, the photoelectron conversion rate has the form of $\lambda_t(r,x_t,\mu_t)$ for $(t,\vec{r})\in[0,\infty) \times \mathcal{R}$ with μ_t an appropriately scaled version of $I_0(t)$, and x is the vector obtained by adjoining x_m and x_p, and H is obtained from H_m and H_p in an obvious way.

The problem of optical tracking is to follow the position of maximum light intensity at time t in terms of both photoelectron conversions observed on $[0,t) \times \mathcal{R}$ and observations of the plant state x_p obtained with sensors located at the tracking devices. These latter observations are modeled according to (1b) so as to account for sensor noise. Except for the finiteness of $\mathcal{R}$, this problem is identical to control problem studied above when photoelectron conversions are identified as space-time points. An approximation that appears reasonable when the beam is small and the tracking errors are small (i.e. fine tracking mode rather than an acquisition mode) compared to the size of the photoemissive surface is to replace $\mathcal{R}$ by R^2. With this approximation, the optical tracking problem is solved by the result in Theorem 2.

It is important to note that according to Remark 3 and Theorem 2, the design of the tracking controller does not depend in any way upon the source or nature of randomness in $I_0(t)$. Thus, for example, the same design is obtained if $I_0(t)$ is random due to atmospheric turbulence or modulation by an information-bearing signal or a combination of these.

The upper and lower bounds of Sections IV and V provide a measure of the performance for the optical position-sensing and tracking system derived from Theorem 2. From Remark 6, the upper and lower bounds merge when HS^*H' is small compared to the beam spread as measured by R. It is evident that the estimation and control lower bounds derived as above for observations of each photoelectron conversion are also lower bounds for both optimal and suboptimal trackers that employ observations obtained by temporal or spatial averaging as would be obtained using photon counting and a quadrant photomultiplier.

We mention also that A. Segall in [11] has applied the models of [1] and [3] to study computer communication networks. The upper and lower bounds on performance that we have derived can be applied in this context as well.

REFERENCES

1. D. L. Snyder and P. M. Fishman, "How to track a swarm of fireflies by observing their flashes", *IEEE Trans. on Info. Theory*, Vol. IT-21, No. 6, pp. 692-695, Nov. 1975.

2. D. L. Snyder, I. B. Rhodes, and E. V. Hoversten, "An exact estimator-controller solution to a stochastic optimal control problem with point process observations", Proc. Sixth Symposium on Nonlinear Estimation and Its Applications, San Diego, Calif., Sept. 1975.

3. D. L. Snyder, I. B. Rhodes, and E. V. Hoversten, "A separation theorem for stochastic control problems with point process observations", *Automatica*, Vol. 13, No. 1, Jan. 1977, to appear.

4. M. V. Vaca, "A measure transformation approach to estimation and decision for observations derived from martingales", D.Sc. dissertation, Sever Institute of Washington University, St. Louis, Mo., May 1975.

5. M. V. Vaca and D. L. Snyder, "Estimation and decision for observations derived from martingales", submitted to IEEE Trans. on Info. Theory.

6. D. L. Snyder, *Random Point Processes*, New York: Wiley, 1975.

7. I. B. Rhodes and A. S. Gilman, "Cone-bounded nonlinearities and mean-square bounds - estimation lower bound", *IEEE Trans. on Automatic Control*, Vol. AC-20, No. 5, pp. 632-642, Oct. 1975.

8. K. J. Åström, *Introduction to Stochastic Control Theory*, New York: Academic Press, Section 8.7, 1970.

9. A. S. Gilman and I. B. Rhodes, "Cone-bounded nonlinearities and mean-square bounds - quadratic regulation bounds", *IEEE Trans. on Automatic Control*, Vo. AC-21, No. 4, pp. 472-483, Aug. 1976.

10. I. B. Rhodes, "A tutorial introduction to estimation and filtering", *IEEE Trans. on Automatic Control*, Vol. AC-16, No. 6, pp. 688-706, Dec. 1971.

11. A. Segall, "Centralized and distributed control schemes for Gauss-Poisson processes", Report ESL-P-669, Electronic Systems Lab., M.I.T., Cambridge, Mass., 1976.

12. I. B. Rhodes and D. L. Snyder, "Estimation and control performance for space-time point-process observations", Report SSM 7610, Department of Systems Science and Mathematics, Washington University, Sept. 1976.

APPENDIX

Lemma. Let W_1 and W_2 be positive definite matrices, and let $\gamma \in [0,1]$. Then

$$[\gamma W_1 + (1-\gamma)W_2] \leq \gamma W_1^{-1} + (1-\gamma)W_2^{-1} \tag{A1}$$

i.e. W^{-1} is convex in a matrix sense. Furthermore, we have

$$E[W^{-1}] \geq (E[W])^{-1}$$

(c.f. Jensen's inequality)

Proof. See [12].

TOWARD A GENERAL THEORY OF CONTROL CANONICAL STRUCTURE[*]

by DAVID L. RUSSELL
Department of Mathematics, University of Wisconsin - Madison

1. Outline of the Theory Let A generate a semigroup, $S(t)$, of bounded operators in the Hilbert space H and consider the control system

$$\dot{x} = Ax + bu, \quad x, b \in H, \quad u \in E^1. \tag{1.1}$$

The solution of (1.1) with $x(0) = x_0 \in H$ is defined by

$$x(t) = S(t)x_0 + \int_0^t S(t-s)\, b u(s)\, ds \tag{1.2}$$

for $t \geq 0$. In order to investigate the effect of the control u we take $x_0 = 0$ and so have

$$x(t) = \int_0^t S(t-s)\, b u(s)\, ds. \tag{1.3}$$

To the same end we make the following

Assumptions on A. The eigenvalues of A are distinct, forming a sequence $\{\lambda_k \mid k = 1, 2, 3, \ldots\}$ and the corresponding eigenvectors, ϕ_k, $k = 1, 2, 3, \ldots$, form a basis for H. Each element $x \in H$ has the convergent representation

$$x = \sum_{k=1}^{\infty} \xi_k \phi_k, \quad \xi_k = (x, \psi_k), \tag{1.4}$$

where ψ_k is the eigenvector of A^* corresponding to the eigenvalue $\bar{\lambda}_k$.

Taking the inner product of (1.1) with ψ_k for

$$x(t) = \sum_{k=1}^{\infty} \xi_k(t)\, \phi_k, \quad b = \sum_{k=1}^{\infty} \beta_k \phi_k,$$

we have

$$\dot{\xi}_k = \lambda_k \xi_k + \beta_k u, \quad k = 1, 2, 3, \ldots. \tag{1.5}$$

The equation (1.3) is then replaced by

$$\xi_k(t) = \beta_k \int_0^t e^{\lambda_k (t-s)} u(s)\, ds, \quad k = 1, 2, 3, \ldots \tag{1.6}$$

Consider now the

Control Problem. Let $y \in H$ and, for $T > 0$, let it be required to find a control $u \in L^2[0,T]$ such that (cf. (1.3))

$$y = x(T) = \int_0^T S(T-t)\, b\, u(t)\, dt. \tag{1.7}$$

Letting the coefficients of the expansion of y in terms of the ϕ_k (cf. (1.4)) be η_k, $k = 1, 2, 3, \ldots$, we see from (1.6) that we should have

---

[*] Supported in part by the Office of Naval Research under Contract NR041-404

$$\eta_k = \beta_k \int_0^T e^{\lambda_k(T-t)} u(t)\,dt\,, \quad k = 1, 2, 3, \ldots \tag{1.8}$$

The equations (1.8) constitute the basic moment problem associated with our control problem. We will make the following

Controllability Assumption. All $\beta_k \neq 0$ and there exists T, $0 < T \leq \infty$, independent of k, and "biorthogonal elements" $\tilde{q}_k \in L^2[0,T]$ such that

$$\int_0^T e^{\lambda_k(T-t)} \tilde{q}_\ell(t)\,dt = \delta_\ell^k = \begin{cases} 1, & k = \ell \\ 0, & k \neq \ell \end{cases}$$

If this assumption is satisfied, (1.8) has the formal solution

$$u(t) = \sum_{k=1}^{\infty} \frac{\eta_k}{\beta_k} \tilde{q}_k(t)\quad ,$$

which is a bona fide solution just in case the series converges in $L^2[0,T]$.

Generally the $\tilde{q}_k$ are not unique and the theory that we will develop requires that they should be. Our program for achieving uniqueness is the following. We let

$$\tau = \inf\{T > 0 \mid \text{the Controllability Assumption is satisfied for } T\}.$$

Consider then the interval $[0, \tau]$ (which might have zero, positive, or infinite length) and let p_k, $k = 1, 2, 3, \ldots$ denote the restrictions to $[0, \tau]$ of the functions

$$e^{\lambda_k(\tau - t)}\quad , \quad k = 1, 2, 3, \ldots .$$

We admit the possibility that the specification of p_k may include some or all of the derivatives of these functions as well as their values. We make the

Uniqueness Assumption. There exist Hilbert spaces U, V associated with the interval $[0, \tau]$ such that

(i) the elements p_k, $k = 1, 2, 3, \ldots$, (whose specification may involve derivatives of $e^{\lambda_k(\tau - t)}$) form a basis for V;

(ii) U is algebraically and topologically isomorphic to $V^* = V$ and contains a basis q_k dual to the basis p_k for V, i.e.

$$\langle p_k, q_\ell \rangle = \delta_\ell^k\ ;$$

(iii) each element $v \in V$ has the unique development

$$v = \sum_{k=1}^{\infty} v_k p_k\,, \qquad v_k = \langle v, q_k \rangle$$

and each element $u \in U$ has the unique development

$$u = \sum_{k=1}^{\infty} u_k q_k \quad , \qquad u_k = \langle p_k , u \rangle$$

We now replace the equations (1.8) by

$$\eta_k = \beta_k \langle p_k , u \rangle \ , \quad u \in U \qquad k = 1, 2, 3, \ldots \tag{1.9}$$

thereby obtaining an abstract generalization of the system (1.8). For this system we have uniqueness of the formal solution

$$u = \sum_{k=1}^{\infty} \frac{\eta_k}{\beta_k} q_k \ . \tag{1.10}$$

In application it will still be possible to interpret the η_k in (1.9) as terminal values of solutions of (1.5) corresponding to certain " distribution valued" controls.

The program for reducing the control system (1.5), (1.9) to control canonical form is begun by using the control to state map described by (1.9) and its inverse, the state to control, or "steering" map, described by (1.10). How this is carried out in detail varies with the individual application, but we can proceed in a fairly general manner here to give the basic ideas.

Consider the change of variable

$$\xi_k(t) = \beta_k \langle p_k , \tilde{\xi}(t, \cdot) \rangle \ , \quad k = 1, 2, 3, \ldots, \tag{1.11}$$

where $\tilde{\xi}(t, \cdot)$ is, for each $t \geq 0$, an element of the space U. If we denote the abstract control operator (1.9) by C, so that (1.9) becomes $\{\eta_k\} = Cu$ then (1.11) becomes

$$\{ \xi_k(t) \} = C \tilde{\xi}(t, \cdot) \tag{1.12}$$

The existence and uniqueness of the biorthogonal elements $q_k \in U$ enables one to define the formal inverse for C, as we have done in (1.10). That equation can be rewritten as

$$u = C^{-1} \{ \eta_k \}$$

and, in the same way, the inverse relationship corresponding to (1.12) is

$$\tilde{\xi}(t, \cdot) = C^{-1} \{ \xi_k(t) \} = \sum_{k=1}^{\infty} \frac{\xi_k(t)}{\beta_k} q_k(\cdot) \ . \tag{1.13}$$

Still proceeding formally (1.12) carries the equations (1.5) into a new system

$$\frac{\partial\tilde{\zeta}(t,\cdot)}{\partial t} = (\text{from } (1.13), (1.5)) = \sum_{k=1}^{\infty} \frac{\lambda_k \xi_k(t) + \beta_k u(t)}{\beta_k} \; q_k(\cdot)$$

$$= (\text{ from } (1.11)) = \sum_{k=1}^{\infty} \lambda_k \langle p_k, \tilde{\zeta}(t,\cdot)\rangle \, q_k(\cdot) + \left(\sum_{k=1}^{\infty} q_k(\cdot)\right) u(t) \; . \qquad (1.14)$$

Writing (1.14) symbolically as

$$\tilde{\zeta}_t = \tilde{A}\,\tilde{\zeta} + \tilde{b}\,u \qquad (1.15)$$

we have the transformed control system in U. One should note that in many cases, however $\tilde{b}$ will not be a vector in U. In fact, since, formally at least,

$$1 = \langle p_k, \delta_{(\tau)} \rangle \; ,$$

where $\delta_{(\tau)}$ denotes the delta function with support $\{\tau\}$, there is reason to suspect that $\sum_{k=1}^{\infty} q_k(\cdot)$ represents this distribution in some sense.

The manner in which the above change of variable transforms the eigenvectors of the system is worthy of note. Each sequence

$$\{\delta_\ell^k \mid \ell = 1, 2, 3, \ldots, \quad \delta_\ell^k = 1 \, , \; \ell = k \; , \quad \delta_\ell^k = 0 \, , \; \ell \neq k\}$$

is an eigenvector for the system (1.5). But

$$C^{-1}\,\delta_\ell^k = (\text{from } (1.13)) = \sum_{\ell=1}^{\infty} \frac{\delta_\ell^k}{\beta_\ell} \, q_\ell(\cdot) = \frac{1}{\beta_k} \, q_k(\cdot) \; .$$

Thus the biorthogonal elements $q_k \in U$ corresponding to the exponentials p_k are eigenvectors for the operator $\tilde{A}$ in (1.15).

The final step involves further transformation of (1.14), (1.15). Let $\Phi : V \rightarrow U$ be defined on the exponentials $p_k(\tau - \cdot)$ (recall $p_k(t) = \exp(\lambda_k(\tau - t))$) by

$$\gamma_k \, \Phi \, p_k(\tau - \cdot) = q_k \; , \quad k = 1, 2, 3, \ldots, \qquad (1.16)$$

with the non-zero γ_k yet to be determined. Again proceeding formally, (1.14) becomes, under the transformation $\tilde{\zeta}(t,\cdot) = \Phi\,\zeta(t,\cdot)$ the equation

$$\frac{\partial\zeta(t,\cdot)}{\partial t} = \sum_{k=1}^{\infty} \gamma_k \lambda_k \langle p_k, \tilde{\zeta}(t,\cdot)\rangle \, p_k(\tau - \cdot) + \left(\sum_{k=1}^{\infty} \gamma_k p_k(\tau - \cdot)\right) u(t) \qquad (1.17)$$

Now, since

$$\gamma_k \langle p_k, \tilde{\zeta}(t, \cdot) \rangle = \langle p_k, \sum_{\ell=1}^{\infty} \frac{\xi_\ell(t)}{\beta_\ell} q_\ell \rangle = \gamma_k \frac{\xi_k(t)}{\beta_k}$$

and

$$\langle \zeta(t, \cdot), q_k(\tau - \cdot) \rangle = \langle \sum_{\ell=1}^{\infty} \frac{\gamma_\ell \xi_\ell(t)}{\beta_\ell} p_\ell(\tau - \cdot), q_k(\tau - \cdot) \rangle = \frac{\gamma_k \xi_k(t)}{\beta_k}$$

We see that (1.17) becomes

$$\frac{\partial \zeta(t, \cdot)}{\partial t} = \sum_{k=1}^{\infty} \lambda_k \langle \zeta(t, \cdot), q_k(\tau - \cdot) \rangle \, p_k(\tau - \cdot) + \left(\sum_{k=1}^{\infty} \gamma_k p_k(\tau - \cdot) \right) u(t). \tag{1.18}$$

Observe now that the exponential functions $p_k(\tau - s) = \exp(\lambda_k s)$ have become the eigenfunctions of the homogeneous system corresponding to the eigenvalue λ_k, since

$$\sum_{\ell=1}^{\infty} \lambda_\ell \langle p_k(\tau - \cdot), q_\ell(\tau - \cdot) \rangle \, p_\ell(\tau - \cdot) = \lambda_k p_k(\tau - \cdot) .$$

This suggests that with $\zeta(t, s)$ redefined as $\zeta(t + s)$ the evolution of $\zeta(t + s)$ can be described by a "scalar" operator of some sort; a scalar differential equation of finite or even infinite order, or perhaps a scalar functional differential equation as in [3], [4].

The choice of the γ_k in (1.16) is arbitrary in principle but in practice is determined by the desirability of having $u(t)$ appear in the final form of (1.18) in as simple a manner as possible. For the hyperbolic system studied in [3], [4] where $\tau = 2$, the series $\sum_{k=1}^{\infty} q_k(\cdot)$ in (1.14) represents $\delta_{(2)}$ and the γ_k are determined by the condition that $\sum_{k=1}^{\infty} \gamma_k p_k(2 - \cdot)$ should still represent $\delta_{(2)}$. In other applications the details are slightly different but the principle remains the same. In the final reduced system the fact that $u(t)$ appears in a simple manner and the fact that the homogeneous part of (1.18) is a scalar operator allows the effect of linear state feedback to be examined to advantage.

2. The Finite Dimensional Case Revisited

In order to establish that the schema of Section 1 really does correspond to reduction to control canonical form as carried out, for example, in [1], [2], [5], let us consider the finite dimensional system

$$\frac{d\zeta_k}{dt} = \lambda_k \xi_k + \beta_k u, \qquad k = 1, 2, \ldots, n, \tag{2.1}$$

where the λ_k are distinct complex numbers. (We remark that the schema of Section 1 is easily modified to allow multiplicity of eigenvalues but space does not permit presentation of that material here.)

In this case the minimal control interval has length zero. The $p_k(t)$ are then the functions $e^{\lambda_k(0-t)} = e^{-\lambda_k t}$, $k = 1, 2, \ldots, n$. We define the space V to be the space of equivalence classes of $n-1$ times differentiable functions with inner product and norm

$$(\hat{f}, f)_V = \sum_{k=0}^{n-1} \hat{f}^{(k)}(0)\, \overline{f}^{(k)}(0), \quad \|f\|_V = \left(\sum_{k=0}^{n-1} |\hat{f}^{(k)}(0)|^2 \right)^{\frac{1}{2}}$$

Members $f, \tilde{f}$ of the same equivalence class satisfy $\|f - \tilde{f}\|_V = 0$. The dual space U consists of distributions

$$u = u^1 \delta_{(0)} + u^2(-\delta'_{(0)}) + \ldots + u^n\left((-1)^{n-1} \delta^{(n-1)}_{(0)}\right)$$

$\delta_{(0)}, \delta'_{(0)}, \ldots, \delta^{(n-1)}_{(0)}$ denoting the delta function with support $\{0\}$ and its successive derivatives. The duality between U and V is expressed by

$$\langle f, u \rangle = \sum_{k=1}^{n} f^{(k-1)}(0)\, u^k\, (-1)^{(k-1)}$$

The components of each function $f \in V$ with respect to the basis $\delta_{(0)}, -\delta'_{(0)}, \ldots, (-1)^{n-1} \delta^{n-1}_{(0)}$ of U are just $f(0), -f'(0), \ldots, (-1)^{(n-1)} f^{(n-1)}(0)$. The components of $e^{-\lambda_k t}$ are therefore $(1, \lambda_k, \ldots, \lambda_k^{n-1})$. Forming the Vandermonde matrix $P = (p^i_j) = (\lambda_i^{j-1})$, P^{-1} exists and we shall call it $Q = (q^i_j)$. The relationship $PQ = I$ shows that the elements

$$q_k = q^1_k \delta_{(0)} + q^2_k(-\delta'_{(0)}) + \ldots + q^n_k\left((-1)^{n-1} \delta^{(n-1)}_{(0)}\right) \quad k = 1, 2, \ldots, n$$

form the basis for U dual to (i.e., biorthogonal to) the basis $e^{-\lambda_k t} \equiv (1, \lambda_k, \ldots, \lambda_k^{n-1})$, $k = 1, 2, \ldots, n$, for V.

Carrying out the coordinate change (1.11) of Section 1, (2.1) becomes (cf. (1.14))

$$\frac{d}{dt}\left(\sum_{\ell=1}^{n} \tilde{\xi}^{\ell}(t)(-1)^{\ell-1} \delta^{(\ell-1)}_{(0)} \right) = \sum_{k=1}^{n} \left(\lambda_k \langle p_k, \sum_{\ell=1}^{n} \tilde{\xi}^{\ell}(t)(-1)^{\ell-1} \delta^{\ell-1}_{(0)} \rangle \right) q_k + \left(\sum_{k=1}^{n} q_k \right) u(t) \qquad (2.2)$$

Now

$$\sum_{k=1}^{n} q_k = (\delta_{(0)}, -\delta'_{(0)}, \ldots, (-1)^{n-1} \delta^{(n-1)}_{(0)}) Q p_1 = \delta_{(0)} .$$

On the other hand

$$\sum_{k=1}^{n} \left(\lambda_k \langle p_k , \sum_{\ell=1}^{n} \tilde{\zeta}^{\ell}(t)\,(-1)^{n-1}\,\delta_{(0)}^{\ell-1} \rangle \right) q_k$$

$$= \sum_{k=1}^{n} \left(\lambda_k \sum_{\ell=1}^{n} p_k^{\ell}\, \tilde{\zeta}^{\ell}(t) \right) q_k = \sum_{k=1}^{n} \left(\sum_{\ell=1}^{n} \lambda_k^{\ell}\, \tilde{\zeta}^{\ell}(t) \right) q_k$$

$$= \sum_{\ell=1}^{n} \tilde{\zeta}^{\ell}(t) \left(\sum_{k=1}^{n} \lambda_k^{\ell}\, q_k \right) .$$

For $\ell = 1, 2, \ldots, n-1$,

$$\sum_{k=1}^{n} \lambda_k^{\ell}\, q_k = \left(\delta_{(0)}, -\delta'_{(0)}, \ldots, (-1)^{n-1}\, \delta_{(0)}^{(n-1)} \right) Q p_{\ell+1} = (-1)^{\ell}\, \delta_{(0)}^{(\ell)}$$

To treat the case $\ell = n$ we note that

$$a(\lambda) = \lambda_k^n + a_n \lambda_k^{n-1} + \ldots + a_2 \lambda_k + a_1 = 0 , \quad k = 1, 2, \ldots, n .$$

where $a(\lambda)$ is the unique (n-1) st degree monic polynomial with zeros λ_1, $\lambda_2, \ldots, \lambda_n$. Thus

$$\sum_{k=1}^{n} \lambda_k^{n}\, q_k = - \sum_{k=1}^{n} \left(\sum_{j=1}^{n} a_j \lambda_k^{j-1} \right) q_k =$$

$$- \sum_{j=1}^{n} a_j \left(\sum_{k=1}^{n} \lambda_k^{j-1}\, q_k \right) = - \sum_{j=1}^{n} a_j (-1)^{j-1}\, \delta_{(0)}^{(j-1)} .$$

Hence, in terms of the basis $\delta_{(0)}, -\delta'_{(0)}, \ldots, (-1)^{n-1}\delta_{(0)}^{(n-1)}$ our system takes the form

$$\frac{d\tilde{\zeta}^1}{dt} = -a_1 \tilde{\zeta}^n + u(t) , \qquad \frac{d\tilde{\zeta}^k}{dt} = \tilde{\zeta}^{k-1} - a_k \tilde{\zeta}^n, \quad k = 2, 3, \ldots, n \qquad (2.3)$$

This will be recognized as the control normal form (cf. [3]), the usual "way station" on the route to the control canonical form.

Continuing, the map Φ defined by (1.16) transforms (1.14) into (1.18). Electing to describe $\zeta(t, s)$ in terms of the basis $1, s, s^2/2, \ldots, s^{n-1}/(n-1)!$ for V, (1.18) becomes

$$\frac{d}{dt} \sum_{\ell=1}^{n} \left(\zeta^{\ell}(t)\, s^{\ell-1}/(\ell-1)! \right) = \sum_{k=1}^{n} \lambda_k \langle \sum_{\ell=1}^{n} \zeta^{\ell}(t) \frac{s^{\ell-1}}{(\ell-1)!} , q_{\ell}(-s) \rangle\, p_k(-s)$$

$$+ \left(\sum_{k=1}^{n} \gamma_k\, p_k(-s) \right) u(t) \qquad (2.4)$$

We note now that the homogeneous part of (2.4) does not depend on the γ_k. Its structure may therefore be examined independent of the choice of the γ_k.

Multiplying (2.4) (via the form $\langle\ ,\ \rangle$) by $\delta_{(0)}^{j-1}$, $j = 1, 2, \ldots, n-1$ and noting that $\langle \frac{s^{\ell-1}}{(\ell-1)!} , \delta_{(0)}^{j-1} \rangle = \delta_{\ell}^{j}$, we obtain (taking $u = 0$ so that we are working with the homogeneous equation)

$$\frac{d\zeta^j(t)}{dt} = \sum_{k=1}^{n} \lambda_k \left(\sum_{\ell=1}^{n} \zeta^\ell(t)\, q_k^\ell \right) \langle\, p_k(-s),\ \delta_{(0)}^{j-1} \,\rangle$$

$$= \sum_{k=1}^{n} \lambda_k^j \sum_{\ell=1}^{n} \left(\zeta^\ell(t)\, q_k^\ell \right) = \sum_{\ell=1}^{n} \left(\sum_{k=1}^{n} \lambda_k^j\, q_k^\ell \right) \zeta^\ell(t)$$

$$= \sum_{\ell=1}^{n} \left(\left(\zeta^1(t),\ \zeta^2(t),\ \ldots,\ \zeta^n(t) \right) Q\, p_{j+1} \right) = \zeta^{j+1}(t) \qquad (2.5)$$

For $j = n$ the calculation proceeds in the same way but one notes that

$$\sum_{\ell=1}^{n} \left(\sum_{k=1}^{n} \lambda_k^n\, q_k^\ell \right) \zeta^\ell(t) = -\sum_{\ell=1}^{n} \left(\sum_{k=1}^{n} \left(\sum_{m=1}^{n} a_m \lambda_k^{m-1} \right) q_k^\ell \right) \zeta^\ell(t)$$

$$= -\sum_{m=1}^{n} a_m\, \zeta^m(t)\,. \qquad (2.6)$$

From (2.5), (2.6) we see that in order to achieve the accepted control canonical form

$$\frac{d\zeta^k}{dt} = \zeta^{k+1}, \quad k = 1, 2, \ldots, n-1, \qquad \frac{d\zeta^n}{dt} = -\sum_{k=1}^{n} a_k\, \zeta^k + \alpha u(t) \qquad (2.7)$$

it is only necessary to require that

$$\sum_{k=1}^{n} \gamma_k\, p_k(-s) = (\text{in } V) = \sum_{k=1}^{n} \gamma_k \left(1,\ \lambda_k,\ \ldots,\ \lambda_k^{n-2},\ \lambda_k^{n-1} \right) = (0,\ 0,\ \ldots,\ 0,\ \alpha)$$

One may verify readily that this is equivalent to

$$\gamma_k = \alpha\, q_k^n\,, \quad k = 1,\ 2, \ldots, n\ .$$

The usual choice, of course, is $\alpha = 1$.

3. Remarks on Control Canonical Form for the Heat Equation

Infinite dimensional generalizations of the control canonical form described in Section 2 were originally explored by this author in [3], [4]. For simple hyperbolic systems the canonical equations were shown to be neutral functional equations with a delay equal to the minimal control time. In [4] some spectral assignement results for the heat equation were derived indirectly from the corresponding results for the wave equation without presenting a canonical form for the heat equation itself. Our goal in this section will be to indicate that the canonical form for the heat equation takes the form of a linear differential equation of infinite order. Our treatment is necessarily very brief.

The equation

$$\frac{\partial w}{\partial t} = \frac{\partial^2 w}{\partial x^2} + r(x)\, w + b(x)\, u(t) \qquad (3.1)$$

with boundary conditions

$$w(0,t) \equiv 0, \quad a_1 w(1,t) + b_1 \frac{\partial w}{\partial x}(1,t) \equiv 0 . \tag{3.2}$$

Such a system is equivalent to the equation

$$\dot{\xi}_k = \lambda_k \xi_k + \beta_k u(t), \quad k = 1, 2, 3, \ldots, \tag{3.3}$$

where the eigenvalues λ_k have the form

$$\lambda_k = -(k - \tfrac{1}{2})^2 \pi^2 + \mathcal{O}(\tfrac{1}{k}), \quad k \to \infty$$

and correspond to orthonormal eigenfunctions $\phi_k \in L^2[0,1]$. We have

$$w(\cdot, t) = \sum_{k=1}^{\infty} \xi_k(t) \phi_k , \quad b = \sum_{k=1}^{\infty} \beta_k \phi_k .$$

If we consider only a finite number of the equation (3.3), say $k = 1, 2, \ldots, n$, the canonical form for that truncated system is (2.7) of Section 2, which is equivalent to the n-th order scalar equation

$$a(D)\zeta = \alpha u(t), \quad a(D) = D^n + a_n D^{n-1} + \ldots + a_2 D + a_1 . \tag{3.3}$$

We have noted in Section 2 that the usual choice is $\alpha = 1$. But the choice

$$\alpha = \prod_{k=1}^{n} (-\lambda_k) \quad \text{corresponding to} \quad \gamma_k = \prod_{k=1}^{n} (-\lambda_k)\, q_k^n$$

is not without merit. Assuming no $\lambda_k = 0$, division of (3.3) by α gives

$$\prod_{k=1}^{n} \left(1 - \frac{D}{\lambda_k}\right) \zeta = u(t) . \tag{3.4}$$

The significance of this choice lies in what happens as $n \to \infty$. The infinite product

$$q(\lambda) = \prod_{k=1}^{\infty} \left(1 - \frac{\lambda}{\lambda_k}\right)$$

converges for all complex λ to an entire function, $q(\lambda)$, of order $\frac{1}{2}$. Expanding in power series about $\lambda = 0$

$$q(\lambda) = \sum_{k=0}^{\infty} \hat{q}_k \lambda^k = \sum_{k=0}^{\infty} q_k (-i\lambda)^k ,$$

$q(\lambda)$ may be interpreted as the Fourier transform of the distribution

$$\sum_{k=0}^{\infty} q_k \delta^{(k)}_{(0)} , \quad \delta^{(k)}_{(0)} = \text{k-th derivative of the Dirac "delta function"} .$$

Thus as $n \to \infty$ the canonical form (3.4) tends to

$$\prod_{k=1}^{\infty} \left(1 - \frac{D}{\lambda_k}\right) \zeta = u(t) \quad \text{or} \quad \sum_{k=0}^{\infty} q_k \zeta^{(k)} = u(t) .$$

This scalar linear differential equation is, in fact, the correct control canonical form for the system (3.1), (3.2) (equivalently (3.3)). We do not have space here to examine the reduction process, i.e. the transformations C and Φ of Section 1, or to give a definition of the spaces U and V. These will appear elsewhere.

REFERENCES

[1] P. Brunovsky : " A classification of linear controllable systems", Kybernetika, 6 (1970), pp. 173-188.

[2] E. B. Lee and L. Markus : " Foundations of Optimal Control Theory", Wiley, New York, 1967.

[3] D. L. Russell : " Canonical forms and spectral determination for a class of hyperbolic distributed parameter systems", M.R.C. Tech. Summary Report # 1614, Math. Res. Ctr., University of Wisconsin, Madison, February 1976. To appear in J. Math. Anal. Appl.

[4] D. L. Russell : " Differential-delay equations as canonical forms for controlled hyperbolic systems with applications to spectral assignment", in Proc. Conf. Contr. Thy. of Distr. Parameter Syst., Nav. Surf. Weap. Ctr., White Oak, Md., May, 1976.

[5] W. M. Wonham : " Linear Multivariable Control", Springer-Verlag, Berlin-Heidelberg-New York, 1974.

DIFFERENTIAL-DIFFERENCE EQUATIONS IN A HILBERT SPACE

RICHARD DATKO
Georgetown University, Washington, D.C. 20057

Abstract

In this report processes involving time delayed systems are extended from a Euclidean space setting to a Hilbert space setting in such a manner that one can discuss problems involving distributed parameters. We first discuss the general class of system under consideration and mention some of its stability properties. Applications are then given to four qualitative problems often associated with delay differential equations in Euclidean space. Because of space limitations proofs will be omitted and the exposition will be restricted to autonomous systems, although all but one of the problems mentioned has a nonautonomous extension.

1. THE BASIC SYSTEM

Notation. $R^+ = [0,\infty)$ will denote the nonnegative half line. H, H_1, H_2 and $\mathcal{H}$ will denote specific real Hilbert spaces and $\mathcal{H}_1$ an arbitrary Hilbert space over the complex field. H^C, H_1^C, H_2^C or $\mathcal{H}^C$ will denote the complex extensions of H, H_1, H_2 and $\mathcal{H}$. The norm in any Hilbert space will be denoted by the symbol $|\cdot|$. The set of continuous linear mappings from one Hilbert space $\mathcal{H}_1$ into another Hilbert space $\mathcal{H}_2$ will be denoted by $[\mathcal{H}_1,\mathcal{H}_2]$ if $\mathcal{H}_1 = \mathcal{H}_2$ the set of mappings will be abbreviated to $[\mathcal{H}_1]$. Let $h > 0$ and $I = [-h,0]$, then $L_2[I,\mathcal{H}_1]$ will denote the equivalence classes of measurable mappings from I into $\mathcal{H}_1$ whose norms are square integrable. If I is replaced by R^+ $L_2[R^+,\mathcal{H}_1]$ will denote an analogous set of equivalence classes. Note that both $L_2[I,\mathcal{H}_1]$ and $L_2[R^+,\mathcal{H}_1]$ are Hilbert spaces (see e.g. [7]). The adjoint of any mapping $B \in [\mathcal{H}_1,\mathcal{H}_2]$ will be denoted by B^*, the conjugate transpose of any vector y in $\mathcal{H}_1$ will be denoted by y^* and the inner product of two vectors y_1 and y_2 in $\mathcal{H}_1$ will be denoted by $\langle y_1,y_2\rangle = y_2^* y_1$.

Let H be a fixed real Hilbert space and $L_2[I,H]$ defined as above. We define the Hilbert space $\mathcal{H} = H + L_2[I,H] = \{(x,\phi): x \in H$ and $\phi \in L_2[I,H]\}$. The norm on $\mathcal{H}$ is given by

$$|\hat{\phi}| = |(x,\phi)| = (|x|^2 + |\phi|^2)^{\frac{1}{2}} = \left[\langle x,x\rangle + \int_{-h}^{0} |\phi(\sigma)|^2 d\sigma\right]^{\frac{1}{2}} \tag{1}$$

If $\hat{\phi} = (x,\phi)$ and $\hat{\psi} = (y,\psi)$ are two points in $\mathcal{H}$ their inner product is defined to be

$$\langle\hat{\phi},\hat{\psi}\rangle = \langle x,y\rangle + \int_{-h}^{0} \langle\phi(\sigma),\psi(\sigma)\rangle d\sigma \ . \tag{2}$$

The complex extension, H^C, of H has similar definitions.

A bilinear mapping, Q, on H_1 is said to be positive Hermitian if $Q(x,y) = Q(y,x)$ for all x,y in H_1 and $Q(x,x) > 0$ if $|x| \neq 0$. This property is symbolically denoted by writing $Q > 0$. If Q is Hermitian and $Q(x,x) \geq 0$ for all x in H_1 we write $Q \geq 0$. $Q \leq 0$ implies that $-Q \geq 0$.

Assumptions on certain operators. Assume that A is a closed linear operator mapping a dense set, $\mathcal{D}(A)$, in H into H and that it is the infinitesimal generator of a semi-group, T(t), of class C_o on [H] (see e.g. [10] for definitions and properties of such semi-groups). The mappings $\{A_j\}$, $1 \leq j \leq m$, are in [H] and $f: R^+ \to H$ is a measurable mapping with whatever properties are necessary to make the problem under discussion coherent.

The basic system. Let $0 < h_1 < h_2 < \cdots < h_m = h$ and consider the delay differential equation formally defined on H by the equations

$$\frac{d(x(t))}{dt} = Ax(t) + \sum_{j=1}^{m} A_j x(t-h_j) + f(t) \tag{3a}$$

if $t \geq 0$, by

$$x(t) = \phi(t), \quad \phi \ \varepsilon \ L_2[I,H] \ \text{ if } \ t \ \varepsilon \ [-h,0) \tag{3b}$$

and by

$$x(0) = x_o \ \varepsilon \ H \ . \tag{3c}$$

Due to the assumptions made on the operator A we cannot in general assume that the system (3) has a solution in H. However it is possible to discuss an "integrated" version of (3) which does have a unique solution for all $\hat{\phi} = (x_o,\phi)$ in H. Thus consider the system

$$x(t,\hat{\phi}) = T(t)x_o + \int_0^t T(t-\sigma) \sum_{j=1}^{m} A_j x(\sigma-h_j,\hat{\phi})d\sigma + \int_0^t T(t-\sigma)f(\sigma)d\sigma \tag{4a}$$

if $t \geq 0$ and

$$x(t,\hat{\phi}) = \phi(t) \ \text{ if } \ t \ \varepsilon \ [-h,0) \ . \tag{4b}$$

In [3] the following theorem is proved.

Theorem 1. If f is locally Bochner integrable then for every $\hat{\phi} = (x_o,\phi)$ in H system (4) possesses a unique solution which for $t \geq 0$ is given by the expression

$$x(t,\phi) = S(t)x_o + \sum_{j=1}^{m} \int_{-h_j}^{0} S(t-h_j-\sigma)A_j\phi(\sigma)d\sigma + \int_0^t S(t-\sigma)f(\sigma)d\sigma . \qquad (5)$$

Here $S: R^+ \to [H]$ is strongly continuous and satisfies (i) $|S(t)| = 0$ if $t < 0$ and (ii)

$$S(t) = T(t) + \int_0^t T(t-\sigma) \sum_{j=1}^{m} A_j S(\sigma-h_j)d\sigma$$

if $t \geq 0$.

Definition 1. If in (5) $|f(t)| = 0$ a.e. on R^+ we shall call $x(t,\hat{\phi})$ the homogeneous solution of (4).

Definition 2. Let $x(t,\hat{\phi})$ be the homogeneous solution of (4) for a given $\hat{\phi}$ in H. Let

$$x_t(\phi) = \{x(t+\sigma,\hat{\phi}): t \geq 0 \text{ and } \sigma \in [-h,0]\} .$$

For each $t \geq 0$ define the point $\hat{x}_t(\hat{\phi})$ in H by the relation

$$\hat{x}_t(\hat{\phi}) = (x(t,\hat{\phi}), x_t(\hat{\phi})) . \qquad (6)$$

Theorem 2. The dynamical system defined on H by the relation (6) generates a semi-group, $\tau(t)$, of class C_o on H. The domain, α, of $\tau(t)$ consists of those $\hat{\phi}$ in H such that ϕ is absolutely continuous with derivative $\dot{\phi}$ in $L_2[I,H]$, $\phi(0) = x_o \in \mathcal{D}(A)$ and $\dot{\phi}(0) = A\phi(0) + \sum_{j=1}^{m} A_j\phi(-h_j)$.

Definition 3. The homogeneous solutions of (4) are said to be uniformly exponentially stable if there exist constants $M \geq 1$ and $\beta > 0$ such that for all $\hat{\phi}$ in H

$$|\hat{x}_t(\hat{\phi})| \leq Me^{-\beta t}|\hat{\phi}| . \qquad (7)$$

The following theorem and its corollary are the basis for a discussion of the asymptotic properties of system (4).

Theorem 3. The homogeneous solutions of (4) are uniformly exponentially stable if and only if for each $\hat{\phi}$ in H $\int_0^\infty |x(t,\phi)|^2 dt < \infty$.

One application of Theorem 3 is the following Lyapunov type corollary.

Corollary. The homogeneous solutions of (4) are uniformly exponentially stable if and only if there exists a $Q \geq 0$ defined on H and a $\delta > 0$ such that the homogeneous solutions of (4) satisfy the inequality

$$\frac{d}{dt} (Q(\hat{x}_t(\hat{\phi}), \hat{x}_t(\hat{\phi})) \leq -\delta|x(t,\hat{\phi})|^2 \qquad (8)$$

for a dense set of $\hat{\phi}$ in H and all $t \geq 0$.

Notice that for each $t \in R^+$ the right side of (8) is the norm of a point in H while the left side is a functional defined on H. The inequality (8) is the analogue of stability results obtained in the Euclidean setting by Hale [9] and Driver [6].

2. SOME QUALITATIVE PROPERTIES OF THE SYSTEM (4)

(a) The Perron condition. Assume f is essentially bounded on R^+, i.e. there exists $M > 0$ such that $|f(t)| \leq M$ a.e. on R^+. The system (4) is said to satisfy the Perron condition (see e.g. [8]) if for each essentially bounded $f: R^+ \to H$

$$\overline{\lim_{t}} \left|\int_0^t S(t-\sigma)f(\sigma)d\sigma\right| < \infty . \tag{9}$$

Theorem 4. The homogeneous solutions of (4) generated a uniformly exponentially stable system if and only if the Perron condition is satisfied.

In engineering parlance Theorem 4 says the homogeneous solutions of (4) decay exponentially if and only if any bounded input to the nonhomogeneous system implies the output is bounded.

(b) Absolute stability. Let $B \in [H_1,H]$, $K \in [H,H_2]$ and F be a continuous nonlinear mapping from H_2 into H_1 which satisfies an inequality of the type

$$|F(\sigma)| \leq L|\sigma| \tag{10}$$

for all $\sigma \in H_2$.

Consider the nonlinear variant of (4) described for $t \geq 0$ by the equations

$$x(t,\hat{\phi}) = S(t)x_o + \sum_{j=1}^{m} \int_{-h_j}^{0} S(t-s-h_j)A_j\phi(s)ds + \int_0^t S(t-s)BF(\sigma(s))ds \tag{11a}$$

where

$$\sigma(s) = Kx(s,\hat{\phi}) \tag{11b}$$

We are interested in the absolute stability of the system (11). That is under what conditions on S(t), K and B can we guarantee that for all $\hat{\phi} \in H$ $|x(t,\hat{\phi})|$ will tend to zero as t tends to infinity? The criteria given below may appear weak. However it should be recalled that in contrast to the finite dimensional situation we may not in general assume every solution of (11) is differentiable in t. However using a procedure due to V. Barbu we can state (see [1] for an exposition):

Theorem 5. Assume the homogeneous solutions of (4) are uniformly exponentially stable. Let

$$\tilde{K}(i\omega) = \int_{-\infty}^{\infty} e^{i\omega t} KS(t)B\,dt, \quad -\infty < \omega < \infty, \tag{12}$$

and assume there exists a nonnegative bilinear Hermitian mapping Q on H_1^C such that for all $\omega \in (-\infty,\infty)$

$$\tilde{K}(i\omega)^*\tilde{K}(i\omega) - Q\tilde{K}(i\omega) - K(i\omega)^*Q \lesssim 0 , \tag{13}$$

then the system (11) is absolutely stable, i.e. $\lim_{t\to\infty} |x(t,\hat{\phi})| = 0$ for all $\hat{\phi}$ in H.

Remark 1. The inequality (13) is a version of the well known frequency criterion for the determination of absolute stability (see e.g. [1] or [8]).

(c) Periodic behavior.

Theorem 6. Assume the homogeneous solutions of (4) generate a uniformly exponentially stable system. Let $f: [0,\infty) \to H$ be periodic of period ω, $\omega > h$. Then there exists a natural number n_o and a unique solution of (4) which is periodic of period $n_o\omega$, i.e. $x(t+n_o\omega,\hat{\phi}) = x(t,\hat{\phi})$ for all $t \in R^+$. The number n_o is independent of the particular f.

Remark 2. One might expect $n_o = 1$. However as yet we are unable to show this. The method of proof in Theorem 6 involves a fixed point argument and uses the fact, that for f fixed, every solution of (4) tends asymptotically to $\int_0^t S(t-s)f(s)ds$.

(d) A control problem involving L_2 stabilization. Let $D \in [H_1,H]$ and $\mu \in L_2[R^+,H]$. Consider the process

$$x(t,\hat{\phi},\mu) = S(t)x_o + \sum_{j=1}^{m} \int_{-h_j}^{0} S(t-s-h_j)A_j\phi(s)ds$$

$$+ \int_0^t S(t-s)D\mu(s)ds \quad \text{if} \quad t \geq 0 \tag{14a}$$

and

$$x(t,\hat{\phi},\mu) = \phi(t) \quad \text{if} \quad t \in [-h,0) . \tag{14b}$$

The problem is to determine conditions for the existence of a feedback which stabilizes (14). That is to find, if possible, a bounded linear mapping $K:H \to H_1$ such that if $\mu(t) = Kx_t(\hat{\phi})$ the process

$$x(t,\hat{\phi}) = S(t)x_o + \sum_{j=1}^{m} \int_{-h_j}^{0} S(t-s-h_j)A_j\phi(s)ds + \int_{0}^{t} S(t-s)DKx_s(\hat{\phi})ds \tag{15}$$

will be uniformly exponentially stable in H.

Theorem 7. The process (14) can be stabilized if and only if for each $\hat{\phi}$ in H there exists $\mu \in L_2[R^+,H_1]$ such that the corresponding solution of (14) satisfies the condition

$$\int_{0}^{\infty} |x(t,\hat{\phi},\mu)|^2 dt < \infty . \tag{16}$$

Remark 3. In a Euclidean setting condition (16) is equivalent to solving the so called regulator problem over the interval $[0,\infty)$ (see e.g. [2] or [5]).

Remark 4. The practical solution of the regulator problem is difficult in Euclidean spaces. However, of one is only interested in stabilization, it is often not necessary to solve the regulator problem. In fact in a Euclidean setting for h "sufficiently small" we only need to stabilize the following system

$$\frac{dx}{dt} = \left(A + \sum_{j=1}^{m} A_j\right)x(t) + D\mu(t) . \tag{17}$$

Furthermore the feedback for (17) will serve as a feedback for (14) (see e.g. [4]). Thus one line of investigation might be to determine, in a Hilbert space setting, the extent to which stabilization of (17) implies stabilization of (14) with the same feedback.

Example. Consider the "damped" telegrapher's equation with delay given by the expression

$$\frac{\partial^2 v}{\partial t^2}(x,t) = \frac{\partial^2 v}{\partial x^2}(x,t) - \frac{\partial v}{\partial t}(x,t-h) + \sum_{j=1}^{\infty} (\sin n\pi x)fn(t) . \tag{18}$$

Here $0 \leqslant x \leqslant 1$ and $t \geqslant 0$. Assume that

$$v(x,s) = \phi(x,s) = \sum_{n=1}^{\infty} \phi_n(s) \sin n\pi x \tag{19a}$$

$$\frac{\partial v}{\partial t}(x,s) = \psi(x,s) = \sum_{j=1}^{\infty} \psi_n(s) \sin n\pi x \tag{19b}$$

for $s \in [-h,0)$,

$$v(0,t) = v(1,t) = 0 \quad \text{for all} \quad t \in [-h,\infty) \tag{19c}$$

and

$$v(x,0) = \sum_{j=1}^{\infty} x_j \sin j\pi x \tag{19d}$$

$$\frac{\partial v}{\partial t}(x,0) = \sum_{j=1}^{\infty} y_j \sin j\pi x \ . \tag{19e}$$

The sequences $\{\phi_n\}$ and $\{\psi_n\}$ are measurable on $[-h,0]$ and together with $\{x_n\}$ and $\{y_n\}$ satisfy the condition

$$\int_{-h}^{0} \sum_{n=1}^{\infty} [n^2\pi^2(\phi_n(s))^2+(\psi_n(s))^2]ds + \sum_{n=1}^{\infty} (n^2\pi^2x_n^2+y_n^2) < \infty \ . \tag{20}$$

The $\{f_n\}$ are measurable on $[0,\infty)$ and

$$\int_0^{\infty} \sum_{j=1} (f_n(t))^2 dt \leq M < \infty \ . \tag{21}$$

Formally separating variables in (18) and applying (19c) we obtain for $t \geq 0$ the solution of the system in the form

$$v(x,t) = \sum_{n=1}^{\infty} T_n(t) \sin n\pi x \ , \tag{22}$$

where for each n T_n satisfies

$$\ddot{T}_n(t) + \dot{T}_n(t-h) + n^2\pi^2T_n(t) = f_n(t) \tag{23}$$

Thus the problem (18)-(19) can be formally reduced to the equation

$$\ddot{x}(t) = Ax(t) + B\dot{x}(t-h) + f(t) \tag{24}$$

in the real sequence space ℓ_2. In (24) $B = -I$ (I is the identity map) and A is the closed unbounded operator which takes the n^{th} coordinate in ℓ_2 into itself multiplied by $-(n\pi)^2$. Clearly (24) is a model for the theory.

REFERENCES

1. C. Corduneanv, Integral Equations and Stability of Feedback Systems, Academic Press, New York and London, 1973.
2. R. Datko, Unconstrained control problems with quadratic cost, S.I.A.M. J. Cont., 11 (1973), pp. 32-52.
3. R. Datko, Linear autonomous neutral differential-difference equations in a Banach space, to appear.
4. R. Datko, The stabilization of linear functional differential equations, Calculus of Variations and Control Theory, D. L. Russell editor, Academic Press, New York and London, 1976, pp. 353-369.
5. M. C. Delfour, C. McCalla and S. K. Mitter, Stability and the infinite-time quadratic cost problem for linear hereditary differential systems, S.I.A.M. J. Cont., 13 (1975), pp. 48-88.

6. R. D. Driver, Existence and stability of solutions of a delay-differential system, Arch. Rational Mech. Anal., 19 (1962), pp. 401-426.

7. R. E. Edwards, Functional Analysis: Theory and Applications, Holt, Rinehart and Winston, New York, 1965.

8. A. Halanay, Differential Equations: Stability, Oscillations, Time Lags, Academic Press, New York, 1966.

9. J. Hale, Functional Differential Equations, Springer Verlag, New York, 1971.

10. E. Hille and R. S. Phillips, Functional Analysis and Semi-Groups, Colloquium Publications, Amer. Math. Soc., Providence 1957.

QUALITATIVE ANALYSIS OF INTERCONNECTED DYNAMICAL SYSTEMS DESCRIBED ON BANACH SPACES: WELL POSEDNESS, BOUNDEDNESS AND INSTABILITY

ANTHONY N. MICHEL*
Electrical Engineering
Department and Engineering
Research Institute
Iowa State University
Ames, Iowa 50011

RICHARD K. MILLER**
Department of Mathematics
Iowa State University
Ames, Iowa 50011

ABSTRACT

New qualitative results for interconnected dynamical systems described on Banach spaces are established. Items of interest include well posedness, instability, complete instability, and Lagrange stability (i.e., boundedness and ultimate boundedness of solutions). In all cases the objective is to analyze complex systems in terms of lower order subsystems and in terms of the system interconnecting structure.

I. INTRODUCTION

In a recent paper [1] sufficient conditions for uniform asymptotic stability and exponential stability of a class of interconnected systems described by differential equations which are defined on Banach spaces were established. Presently we continue the work initiated in [1] by considering questions of well posedness, Lagrange stability (i.e., boundedness and ultimate boundedness of solutions) and instability of composite dynamical systems described on infinite dimensional spaces.

The reader is referred to [1] for additional references dealing with the qualitative analysis of interconnected systems, which are also called composite systems, large scale systems, decentralized systems, multi loop systems, and so forth.

II. PRELIMINARIES

The present section consists of four parts. First we establish the necessary notation and then we present pertinent facts concerning dynamical systems defined on Banach spaces. In the third part we introduce admissible Lyapunov functions while in the last part we characterize the class of interconnected systems considered herein.

---

* The research of the first author was supported by the National Science Foundation under Grant Eng 75-14093 and by the Engineering Research Institute, Iowa State University.

** The research of the second author was supported by the National Science Foundation under Grant MPS 75-07622.

A. Notation

Let $R = (-\infty,\infty)$ and $R^+=[0,\infty)$. A function $\psi:R^+\to R^+$ is said to belong to class K (i.e., $\psi\in K$) if it is continuous, strictly increasing and if $\psi(0) = 0$. If $\psi\in K$ and $\lim\limits_{r\to\infty}\psi(r) = \infty$, then ψ is said to belong to class KR. Two functions $\psi_1,\psi_2\in K$ are said to be of the same order of magnitude if there exist positive constants k and ℓ such that $k\psi_1(r) \leq \psi_2(r) \leq \ell\psi_1(r)$ for all $r\in R^+$. (For a discussion of the properties of functions of class K see [2].) The notation $\mathcal{O}(r)$ denotes terms of higher order, so that $\lim\limits_{r\to 0^+}\mathcal{O}(r)/r = 0$.

Let $\{R^n;|\cdot|\}$ denote the n-dimensional Euclidean space with norm $|\cdot|$. If $x\in R^n$, then $x^T = (x_1,\dots,x_n)$ denotes the transpose of x, $x> 0$ means $x_i > 0$ and $x\geq 0$ means $x_i\geq 0$, $i = 1,\dots,n$. The transpose of a matrix C is denoted by C^T, $C = [c_{ij}] > 0$ means $c_{ij} > 0$ and $C\geq 0$ means $c_{ij}\geq 0$ for all i and j.

Banach spaces are denoted by X or Z, with appropriate subscripts if necessary and norms on such spaces are denoted by $\|\cdot\|$, with appropriate subscripts as required. $I:X\to X$ will denote the identity transformation and $\|A\|$ denotes the norm of a bounded linear operator $A:D(A)\to Z$, $D(A)\subset X$.

B. Semigroups

Appropriate references for this section include [3] and [4].

Definition 1. Assume C is a subset of a Banach space X. A function $T(t):C\to C$ is called a semigroup on C if $T(t)x$ is continuous in (t,x) on $R^+\times C$, $T(0)x = x$ and $T(t+s)x = T(t)T(s)x$ when t and $s\in R^+$ and $x\in C$. If there is a number $w\in R$ such that $\|T(t)x-T(t)y\|\leq e^{wt}\|x-y\|$ for all $t\in R^+$ and all $x,y\in C$, then $T(t)$ is called a quasicontractive semigroup. On the other hand if $C = X$ and if $T(t)$ is a bounded linear map for each fixed $t\in R^+$, then $T(t)$ is called a C_0-semigroup.

The infinitesimal generator A_s of a semigroup $T(t)$ is defined by

$$A_s(x) = \lim_{t\to 0^+}\frac{T(t)x-x}{t},\ x\in D(A_s)$$

for all x such that the limit exists.

Next, consider a differential equation of the form

$$\dot{x}(t) = A(x(t)),\ x(0) = x_0\ \in D(A). \tag{E}$$

First assume that $A:D(A)\to X$ is linear and $D(A)$ is dense in X. By a strong solution of (E) we mean a function $x:R^+\to D(A)$ such that $\dot{x}(t)$ exists and is continuous on R^+ into X, $x(0) = x_0$ and

(E) is true for all $t \in R^+$. Call (E) well-posed if for each $x_0 \in D(A)$ there is a strong solution $x(t,x_0)$ and this solution depends continuously on (t,x_0). It is known [3,4] that (E) is well posed if and only if $T(t)x_0 = x(t,x_0)$ determines a C_0-semigroup on X. In this case the infinitesimal generator A_s is the closure of A.

If $A:D(A) \to X$ is nonlinear with D(A) dense in C, then in general we must allow A to be multivalued. In this case a strong solution of (E) is a locally absolutely continuous function $x:R^+ \to C$ such that $x(t) \in D(A)$ and $\dot{x}(t) \in A(x(t))$ a.e. on R^+. Nonlinear problem (E) is called well-posed on C if there is a semigroup T(t) on C such that for all $x_0 \in D(A)$, $T(t)x_0$ is a strong solution of (E) and D(A) is dense in C. We say that a nonlinear A generates a semigroup if for all $x \in C$,

$$T(t)x = \lim_{n \to \infty} (x - \frac{t}{n} A(x))^{-n}.$$

Definition 2. A multivalued function A on X is called w-accretive if

$$\|(x_1 - \lambda y_1) - (x_2 - \lambda y_2)\| \geq (1 - \lambda w)\|x_1 - x_2\|$$

for all $\lambda \geq 0$, for all $x_i \in D(A)$ and all $y_i \in A(x_i)$, $i = 1,2$.

C. Admissible Lyapunov Functions

Subsequently T(t) always denotes a semigroup on C with the origin 0 in the interior of C and with a corresponding well posed (E).

Definition 3. A function v is called a Lyapunov function if there is an open neighborhood $U \subset X$ of the origin such that $v:U \to R$ is continuous and such that $v(0) = 0$. For such a function

$$Dv(x) = \limsup_{t \to 0^+} \frac{v(T(t)x) - v(x)}{t}$$

is the upper right Dini derivative along trajectories.

Definition 4. A pair (v,T) is called admissible if v is a Lyapunov function for T(t), if the infinitesimal generator A_s of T(t) is defined on a set $D(A_s)$ dense in C, and if there is a function ∇v defined on $\{U \cap D(A_s)\} \times X$ with values in R such that

(i) $v(y) - v(x) \leq \nabla v(x, y-x) + \Theta(\|y-x\|)$ for all $x,y \in D(A_s)$, and

(ii) $\nabla v(x,h)$ is, for each fixed x, a bounded linear operator in $h \in X$.

We will require the subsequent result.

Theorem 1. If (v,T) is an admissible pair and if there exists a function $\varphi \in K$ such that $\nabla v(x, A_s x) \leq -\varphi_3(\|x\|)$ for all $x \in D(A_s) \cap U$, then $Dv(x) \leq -\varphi_3(\|x\|)$ for all $x \in U$.

D. Interconnected Systems

Consider ℓ isolated subsystems $(\mathcal{S}_i)$ described by equations of the form

$$\dot{z}_i = f_i(z_i), \tag{$\mathcal{S}_i$}$$

$i = 1, \ldots, \ell$, which are assumed to be well posed on Banach spaces Z_i with respective norms $\|\cdot\|_i$. Let $T_i(t)$ be the semigroup generated by f_i. $T_i(t)$ is defined on a set $C_i \subset Z_i$ where the origin is contained in the interior of C_i and $C_i = Z_i$ is allowed. The domain $D(f_i)$ must be dense in C_i. We note that for a given integer i system $(\mathcal{S}_i)$ might represent a system of ordinary differential equations, a system of delay differential equations, or perhaps a system of partial differential equations. In the case where $T_i(t)$ is a nonlinear semigroup we must allow the possibility that f_i is multivalued, i.e., a relation. In this case system $(\mathcal{S}_i)$ should actually be replaced by

$$\dot{z}_i \in f_i(z_i).$$

Next, we interconnect these subsystems to form a system (Σ_i) described by equations of the form

$$\dot{z}_i = f_i(z_i) + g_i(x), \tag{Σ_i}$$

$i = 1, \ldots, \ell$, where the operators g_i, representing the interconnecting structure of (Σ_i) are defined on $D(g_i) \subset X$ and have range in Z_i. The hypervector $x^T = (z_i, \ldots, z_\ell)$ is a point in the product space $X = \times_{i=1}^{\ell} Z_i$ with the norm specified by $\|x\| = (\sum_{i=1}^{\ell} \|z_i\|_i^2)^{1/2}$. Letting $f(x)^T = [f_1(z_1), \ldots, f_\ell(z_\ell)]$, $g(x)^T = [g_1(x), \ldots, g_\ell(x)]$, we can express (Σ_i) equivalently as

$$\dot{x} = f(x) + g(x) \triangleq A(x). \tag{$\mathcal{S}$}$$

We call system $(\mathcal{S})$, which is a special case of system (E), an interconnected system or a composite system with decomposition (Σ_i). Note that $D(f+g) = D(f) \cap D(g) = D(f) \cap D(g_1) \cap \ldots \cap D(g_\ell)$ is the domain for $(\mathcal{S})$. Henceforth, each subsystem $(\mathcal{S}_i)$ as well as composite system $(\mathcal{S})$ is assumed to have the trivial solution $z_i(t) \equiv 0$ and $x(t) \equiv 0$, respectively. In addition, the composite system $(\mathcal{S})$ is assumed to be well posed, i.e., $f+g$ generates a semigroup $T(t)$.

III MAIN RESULTS

The present section consists of four parts. In the first of these we consider *well posedness* of system ($\mathcal{S}$), in the second part we present an *instability* result for ($\mathcal{S}$), in the third part we give a result for *boundedness* and *ultimate boundedness* of solutions of ($\mathcal{S}$), and in the fourth part we use properties of Minkowski matrices to simplify some of our results.

A. Well Posedness

Our well posedness results are as follows.

Theorem 2. Let $A_i:D(A_i)\to Z_i$ be a linear operator with domain $D(A_i)$ dense in Z_i. If $f_i(z_i)\overset{\Delta}{=}A_iz_i$ generates a C_0-semigroup on Z_i for each $i=1,\ldots,\ell$, if $D(g_i)=X$ and if $g_i:X\to Z_i$ is a bounded linear operator for each i, then composite system ($\mathcal{S}$) with decomposition (Σ_i) is *well posed* on X.

Theorem 3. If $f_i(z_i)$ is w_i-accretive and linear on a uniformly convex space Z_i, if $Ra(I+\lambda f_i)=Z_i=\overline{D(f_i)}$ for all λ in an interval $(0,\lambda_0)$, if $D(g_i)=X$ with g_i Lipschitz continuous, and if $g_i(0)=0$, $i=1,\ldots,\ell$, then composite system ($\mathcal{S}$) with decomposition (Σ_i) is *well posed* and generates a quasicontraction on X.

B. Instability

In discussing instability results for composite system ($\mathcal{S}$), we find it convenient to employ the following conventions.

Definition 5. Isolated subsystem ($\mathcal{S}_i$) is said to possess *Property C* if there exists a Lyapunov function v_i such that (v_i,T_i) is admissible in the sense of Definition 4, two functions $\varphi_{i1},\varphi_{i3}\epsilon K$ and real constants σ_i and $m_i>0$ such that

$$\varphi_{i1}(\|z_i\|_i)\leq -v_i(z_i)$$
$$\nabla v_i(z_i,f_{is}(z_i))\leq\sigma_i\varphi_{i3}(\|z_i\|_i)$$

for all $z_i\epsilon D(f_{is})$, where $D(f_{is})$ denotes the domain of the infinitesimal generator of $T_i(t)$, with $\|z_i\|_i<m_i$.

Definition 6. Isolated subsystem ($\mathcal{S}_i$) is said to possess *Property C'* if there exists a Lyapunov function v_i with (v_i,T_i) admissible in the sense of Definition 4, a function $\varphi_{i3}\epsilon K$ and real constants σ_i and $m_i>0$ such that

$$\nabla v_i(z_i,f_{is}(z_i))\leq\sigma_i\varphi_{i3}(\|z_i\|_i)$$

for all $\|z_i\|_i<m_i, z_i\epsilon D(f_{is})$.

Remark 1. If in Definition 5, $\sigma_i<0$, then the trivial solution of subsystem $(\mathcal{S}_i)$ is unstable (in fact, completely unstable). (Refer to [2] for the principal Lyapunov results for dynamical systems described on Banach spaces.)

Theorem 4. Let N be a nonempty subset of $L = \{1,\ldots,\ell\}$. Assume that for composite system $(\mathcal{S})$ with decomposition (Σ_i) the following conditions hold:

(i) if $i \in N$, then $(\mathcal{S}_i)$ possesses Property C and if $i \notin N, i \in L$, then $(\mathcal{S}_i)$ possesses Property C';

(ii) there exist real constants b_{ij} such that the inequalities

$$\nabla v_i(z_i, g_i(x)) \le \varphi_{i3}(\|z_i\|_i)^{1/2} \sum_{j=1}^{\ell} b_{ij}\varphi_{j3}(\|z_j,\|_j)^{1/2}$$

hold for all $x^T = (z_1,\ldots,z_\ell) \in D(f+g)$ and $\|z_i\|_i < m_i$ for all $i \in L$;

(ii) there exist positive constants α_i for all $i \in L$ such that the test matrix $R = [r_{ij}]$ defined by

$$r_{ij} = \begin{cases} \alpha_i(\sigma_i + b_{ii}) & \text{if } i = j \\ \frac{1}{2}(\alpha_i b_{ij} + \alpha_j b_{ji}) & \text{if } i \neq j \end{cases}$$

is negative definite.

Then the trivial solution of composite system $(\mathcal{S})$ is unstable.

Remark 2. If in Theorem 4 $N=L$, then the equilibrium of system $(\mathcal{S})$ is completely unstable.

C. Boundedness and Ultimate Boundedness

Next, we consider the boundedness of solutions of $(\mathcal{S})$.

Definition 7. Isolated subsystem $(\mathcal{S}_i)$ is said to possess Property D if there exists a Lyapunov function v_i defined for all $z_i \in Z_i$ with (v_i, T_i) admissible in the sense of Definition 4, three functions $\varphi_{i1}, \varphi_{i2}, \varphi_{i3} \in KR$, a constant $\sigma_i \in R$, and positive constants M_i and R_i, such that the inequalities

$$\varphi_{i1}(\|z_i\|_i) \le v_i(z_i) \le \varphi_{i2}(\|z_i\|_i)$$

and

$$\nabla v_i(z_i, f_{is}(z_i)) \le \sigma_i \varphi_{i3}(\|z_i\|_i)$$

hold when $z_i \in D(f_{is})$ and $\|z_i\|_i \ge R_i$, and if $|v_i(z_i)| \le M_i$ and $|\nabla v_i(z_i, f_{is}(z_i))| \le M_i$ when $\|z_i\|_i \le R_i$ and $z_i \in D(f_{is})$.

Remark 3. If in Definition 7 $\sigma_i < 0$, then the solutions of isolated subsystem $(\mathcal{S}_i)$ are uniformly bounded, in fact, uniformly ultimately bounded.

(Refer to [2] for the principal boundedness results of dynamical systems described on Banach spaces.)

Theorem 5. Assume that for composite system ($\mathcal{S}$) with decomposition (Σ_i) the following conditions hold:

(i) each isolated subsystem ($\mathcal{S}_i$) possesses Property D;

(ii) there exist real constants $b_{ij}, i,j = 1,\ldots,\ell$, such that the inequalities

$$\nabla v_i(z_i, g_i(x)) \leq \varphi_{i3}(\|z_i\|_i)^{1/2} \sum_{j=1}^{\ell} b_{ij}\varphi_{j3}(\|z_j\|_j)^{1/2}$$

hold for all $x^T = (z_1,\ldots,z_\ell) \in D(f+g)$; and

(iii) there exist positive constants $\alpha_i, i=1,\ldots,\ell$, such that the test matrix $R = [r_{ij}]$ defined by

$$r_{ij} = \begin{cases} \alpha_i(\sigma_i + b_{ii}) & \text{if } i = j \\ \frac{1}{2}(\alpha_i b_{ij} + \alpha_j b_{ji}) & \text{if } i \neq j \end{cases}$$

is negative definite.

Then the solutions of system ($\mathcal{S}$) are uniformly bounded, in fact, uniformly ultimately bounded.

D. Application of M-matrices

If in Theorems 4 and 5 $b_{ij} \geq 0$, $i,j=1,\ldots,\ell$, $i \neq j$, as is the case in many applications, then we can use the theory of M-matrices (see [6,7]) to advantage.

Definition 8. An $\ell \times \ell$ matrix $A = [a_{ij}]$, $i,j = 1,\ldots,\ell$, is said to be an M-matrix if $a_{ij} \leq 0$ for all $i \neq j$ and if any one of the following equivalent statements is true:

(i) all principal minors of A are positive;

(ii) all successive principal minors of A are positive;

(iii) there is a vector $x > 0$ such that $Ax > 0$;

(iv) there is a vector $y > 0$ such that $A^T y > 0$;

(v) A is nonsingular and $A^{-1} \geq 0$; and,

(vi) the real parts of all eigenvalues of D are positive.

The applicability of M-matrices to Theorems 4 and 5 lies in the following result, which is a consequence of the preceding statements.

Corollary 1. Let $D = [d_{ij}]$ be an $\ell \times \ell$ matrix with nonpositive off-diagonal elements. Then D is an M-matrix if and only if there exists a matrix $A = \mathrm{diag}[\alpha_1,\ldots,\alpha_\ell]$, $\alpha_i > 0$, $i=1,\ldots,\ell$, such that the matrix $B=AD+D^TA$ is positive definite.

Theorems 4 and 5, Corollary 1 and Definition 8 yield Corollaries 2 and 3. In these results, the test matrix $S=[s_{ij}]$ is specified by

$$s_{ij} = \begin{cases} -(\sigma_i+b_{ii}) \text{ if } i=j \\ -b_{ij} \text{ if } i\neq j \end{cases} . \tag{1}$$

Corollary 2. Assume that for composite system ($\mathcal{S}$) with decomposition (Σ_i) the following is true:

(i) hypotheses (i) and (ii) of Theorem 4 are satisfied with $b_{ij} \geq 0$ for all $i \neq j$; and

(ii) any one of the following equivalent conditions hold:

(a) all successive principal minors of the test matrix $S = [s_{ij}]$ given in (1) are positive;

(b) the real parts of the eigenvalues of S are positive;

(c) there exist positive constants λ_i, $i=1,\ldots,\ell$, such that

$$-(\sigma_i+b_{ii}) - \sum_{\substack{j=1\\ i\neq j}}^{\ell} \frac{\lambda_j}{\lambda_i} b_{ij} > 0, \quad i=1,\ldots,\ell \ ;$$

(d) there exist positive constants $\varkappa_i$, $i=1,\ldots,\ell$, such that

$$-(\sigma_i+b_{ii}) - \sum_{\substack{j=1\\ i\neq j}}^{\ell} \frac{\varkappa_j}{\varkappa_i} b_{ji} > 0, \quad i=1,\ldots,\ell.$$

If $N\neq L$, then the trivial solution of composite system ($\mathcal{S}$) is unstable. If $N=L$, then the trivial solution of ($\mathcal{S}$) is completely unstable.

Corollary 3. Assume that for composite system ($\mathcal{S}$) with decomposition (Σ_i) the following conditions are true:

(i) Hypotheses (i) and (ii) of Theorem 5 are satisfied with $b_{ij} \geq 0$ for all $i \neq j$; and

(ii) any one of the four equivalent conditions in hypothesis (ii) of Corollary 2 is satisfied.

Then the solutions of ($\mathcal{S}$) are uniformly bounded, and in fact, uniformly ultimately bounded.

Remark 4. If in Definition 7 $R_i=0$, $i=1,\ldots,\ell$, then Theorem 5 and Corollary 3 yield sufficient conditions for the uniform asymptotic stability in the large of the trivial solution of composite system ($\mathcal{S}$) (see Theorem 1 and Corollaries 1 and 2 in [1]).

Remark 5. If in Definition 7 $R_i=0$, $i=1,\ldots,\ell$, and if $\varphi_{ij}\epsilon KR$ and are of the same order of magnitude for $i=1,\ldots,\ell$ and $j=1,2,3$, then Theorem 5 and Corollary 3 yield sufficient conditions for the exponential stability in the large of the trivial solution of composite system ($\mathcal{S}$) (see Theorem 2 in [1]).

IV. CONCLUDING REMARKS

Concerning the above results, several comments are in order.

1. The analysis of interconnected dynamical systems (also called composite systems, large scale systems) is accomplished in terms of simpler subsystems and in terms of the system interconnecting structure.

2. The present results are general enough to allow analysis of dynamical systems described by ordinary differential equations, differential-difference equations, functional differential equations, Volterra integro-differential equations, certain classes of partial differential equations, and the like. In addition, the present results allow analysis of hybrid dynamical systems (i.e., systems described by a mixture of equations) in a systematic fashion.

3. Note that Theorems 2 and 3 are stability theorems for composite system ($\mathcal{S}$) in the sense that well posedness is a weak type of stability. Theorem 3 is good enough for many applications, however, it would be interesting to obtain a result of this type which allows a mixed group of quasi-contractive subsystems and linear C_0-subsystems.

4. Existing instability results for interconnected systems described on finite dimensional spaces impose the requirement that all subsystems be unstable (thus, obtaining complete instability results), while Theorem 4 does not make this unreasonable assumption. (Refer to Rasmussen and Michel [1] for references dealing with stability and instability of finite dimensional interconnected systems.)

5. The proof of Theorem 5 does not constitute a trivial modification of the proof of Theorem 1 given in [1] (for uniform asymptotic stability).

6. It is important to observe that in the case of Theorems 4 and 5 (and also in Theorems 1 and 2 given in [1]) the off diagonal elements of the test matrix R can have arbitrary sign. As such, these results are applicable to strongly coupled systems as well as to weakly coupled systems. On the

other hand, Corollaries 2 and 3 (and Corollaries 1 and 2 given in [1], as well as most existing results on finite dimensional composite systems cited in [1]) involve test matrices with off diagonal terms of the same sign. As such, these results constitute weak coupling conditions.

7. It is emphasized that in the case of systems described on infinite dimensional spaces, questions of well posedness are important considerations, frequently giving rise to difficult problems.

REFERENCES

[1] R. D. Rasmussen and A. N. Michel, "Stability of Interconnected Dynamical Systems Described on Banach Spaces", IEEE Transactions on Automatic Control, vol. 21, August 1976, pp. 464-471.

[2] W. Hahn, Stability of Motion. New York: Springer-Verlag, 1967.

[3] A. Pazy, "Semigroups of Linear Operators and Applications to Partial Differential Equations", University of Maryland, Department of Mathematics, Lecture Notes No. 10, 1974.

[4] S. G. Krein, Linear Differential Equations in Banach Spaces, Providence, R. I.: Translations of Mathematical Monographs, vol. 29, Amer. Math. Soc., 1970.

[5] N. Dunford and J. R. Schwarz, Linear Operators, Part I, New York, N. Y.: Interscience Publishers, 1958.

[6] A. Ostrowski, "Determinanten mit überwiegender Hauptdiagonale und die absolute Konvergenz von linearen Iterationprozessen", Comentarii Math. Helvetici, vol. 30, 1956, pp. 175-210.

[7] M. Fiedler and V. Ptak, "On Matrices with Non-positive Off-diagonal Elements and Positive Principal Minors", Czech. Math. J., vol. 12, pp. 382-400, 1962.

ON THE RELATIONSHIP BETWEEN FIRST AND SECOND ORDER CONTROLLABLE SYSTEMS IN BANACH SPACES

ROBERTO TRIGGIANI, Mathematics Department, Iowa State University,
Ames, Iowa 50011

1. Introduction

Consider the abstract control systems

$$\mathbf{S} : \ddot{x} = Ax+Bu \qquad (\mathbf{S}_m : \ddot{x} = Ax+\textstyle\sum_{i=1}^m b_i u_i,\ b_i \in X,\ u_i = \text{ scalar})$$

$$\mathbf{F} : \dot{x} = Ax+Bu \qquad (\mathbf{F}_m : \dot{x} = Ax+\textstyle\sum_{i=1}^m b_i u_i,\ b_i \in X,\ u_i = \text{ scalar})$$

($\mathbf{S}$ and $\mathbf{F}$ stand for second order and first order system, respectively), where both X and U are complex separable Banach spaces and B is a bounded linear operator from U into X. Unless otherwise stated, X is always intended infinite dimensional. $\mathbf{S}_m$ and $\mathbf{F}_m$ refer to the case when dim U = m or more generally dim BU = m, BU = range of B. When m=1, we shall write b instead of b_1. We shall also employ the notation $\mathbf{S}(A,B)$, $\mathbf{S}_m(A,(b_1,..,b_m))$, etc. to emphasize a particular pair (A,B) and $(A,(b_1,..,b_m))$ under consideration. The operator A is assumed throughout to satisfy the following assumption.

H1. A is (closed, linear, with domain D(A) dense in X and range in X and) the infinitesimal generator of a strongly continuous cosine function C(t) of bounded linear operators in X, $-\infty<t<\infty$.

Such assumption is necessary and sufficient for the homogeneous second order-system to be uniformly well posed on the real line and of type $\leq \omega_0$ (F2,I,Thm.5.9,p.91). See e.g. (F2,K2-K4) and the references cited therein for the necessary background on the theory of abstract cosine functions.

A main theme of the present paper is a comparison between controllability properties of $\mathbf{S}$ and $\mathbf{F}$ (for precise definition see §2), which was already carried out in (F1) by Fattorini. He showed the following results (using our terminology):

1) approximate controllability in finite time for $\mathbf{S}$ is always stronger than approximate controllability in finite time for $\mathbf{F}$, while these two properties are equivalent if a certain assumption H2 (below) regarding the spectrum of A holds. The assumption in question is

H2. There exists a simple closed curve Γ entirely contained in $\rho_0(A)$ - the connected component of $\rho(A)$ that contains the half plane $Re\lambda > \omega_0^2$ and such that the origin is contained in the interior of Γ.

2) If H2 is not satisfied, the above equivalence may fail to hold. Example: Let $X = L_2(-\infty,\infty)$; A be the self adjoint operator defined by $(Af)(\xi) = f''(\xi)+rf(\xi)$ with $D(A) = \{f \in X: f'' \in X\}$, ($f''$ understood in the sense of distributions), $r \geq 0$; $b_1(\xi) = e^{-|\xi|}$ and $b_2(\xi) = b_1(\xi+1)$. Then, (i) assumption H2 is not satisfied; (ii) $\mathbf{F}_2(A,(b_1,b_2))$ is approximately controllable in finite time, while (iii) $\mathbf{S}_2(A,(b_1,b_2))$ is not.

On the basis of the above results of Fattorini therefore, one may be led to conjecture that assumption H2 is a necessary condition for the implication: "$\mathbf{F}$ approximately controllable in finite time → $\mathbf{S}$ approximately controllable in finite time" to hold.

As the present paper will show, it turns out however that Fattorini's result is crucially related to his approach, which is based on the intermediary action between $\mathbf{F}$ and $\mathbf{S}$ played by the resolvent operator $R(\lambda,A)$ of A, and it is expressed by

$$R(\lambda^2,A)x = \frac{1}{\lambda}\int_0^\infty e^{-\lambda t}C(t)x\,dt = \int_0^\infty e^{-\lambda^2 t}S(t)x\,dt, \quad \mathrm{Re}\lambda > \omega_0^2,\ x\varepsilon X \tag{1.1}$$

Here $S(t)$ is the semigroup associated with A (see §2).

In the present paper we take a different approach, based in the relationship (see (2.3) below) between the cosine function and the semigroup; (i.e. we work in the t-domain rather than in the λ-domain) and we obtain new results which complement Fattorini's.

2. Comparison between approximate controllability of **S** and **F**

The mild solution of the Cauchy problem associated with the system **S** is, by definition,

$$x_s(t,x_0,x_1,u) = C(t)x_0+\int_0^t C(\tau)d\tau x_1+\int_0^t\left(\int_0^{t-\tau}C(s)ds\right)Bu(\tau)d\tau \tag{2.1}$$

$$\dot{x}_s(t,x_0,x_1,u) = C'(t)x_0+C(t)x_1+\int_0^t C(t-\tau)Bu(\tau)d\tau \qquad t\geq 0 \tag{2.2}$$

which make sense for any x_0,x_1 in X and any Bochner integrable (locally L_1) abstract function $u(\cdot)$. In particular, if $x_0,x_1 \varepsilon D(A)$ and $u(\cdot)$ is C^1, the mild solution is indeed the strict solution (twice strongly continuously differentiable) of **S** (F1). It is a fundamental fact - which will be exploited throughout in the sequel - than an operator A satisfying H1 automatically generates a strongly continuous (C_0) semigroup $S(t)$ of bounded linear operators on X, $t\geq 0$, which may be extended analytically onto the half-plane $\mathrm{Re}\ t > 0$. Moreover, $S(t)$ is given by[1]

$$S(t)x = \frac{1}{\sqrt{\pi t}}\int_0^\infty e^{-\tau^2/4t}\,C(\tau)xd \quad \text{for} \quad t>0,\ x\varepsilon X \tag{2.3}$$

(F2, Eq(5.17),K3). Hence the mild solution of the Cauchy problem associated with the system **F** is by definition

$$x_f(t,x_o,u) = S(t)x_0+\int_0^t S(t-\tau)Bu(\tau)d\tau \quad, t\geq 0 \tag{2.4}$$

for any x_o in X and any Bochner integrable $u(\cdot)$. In particular, if $u(\cdot)$ is Hölder continuous, the mild solution is indeed the strict solution (strongly continuously differentiable) of **F**. (K1,p.491). In what follows, we shall take for both **S** and **F** zero initial conditions. As in Fattorini (F1), we wish to compare controllability properties of **F** and **S** according to the following definitions. Let $K_t(\mathbf{S})$ be the set of attainability from the origin of **S**, i.e. the linear subspace of $X\times X$ consisting of mild solution paris $(x_s(t,0,0,u),\ \dot{x}_s(t,0,0,u))$ when u runs over $L_1((0,t),U)$. Similarly, $K_t(\mathbf{F})$ is the set of attainability from the origin of **F**, i.e. the linear subspace of X consisting of mild solution points $x_f(t,0,u)$ as u runs over $L_1((0,t),U)$.

The space $X\times X$ of all pairs (x_1,x_2) of elements of X is endowed with pointwise operations and with norm $|||(x_1,x_2)||| = \{||x_1||^2+||x_2||^2\}^{\frac{1}{2}}$, so that (K1 p. 164) $(X\times X)^* = X^*\times X^*$ (as Banach spaces). Also $x_1^*(x_1)+x_2^*(x_2)$, $x_i^* \varepsilon X_i$ is the application of a bounded linear functional in $(X\times X)^*$ to $(x_1,x_2) \varepsilon X\times X$. We then say that[2] **S** is approximately controllable in $(0,T), 0<T<\infty$, (respectively, in finite time) in case $\overline{K_T(\mathbf{S})} = X\times X$ (respectively, $\overline{\bigcup_{0<t}K_t(\mathbf{S})} = X\times X$). Similarly, we say that F is approximately controllable in $(0,T)$ in case $\overline{K_T}(F) = X$. It is a

standard fact (deducible from (2.5) below) that, since $S(t)$ is analytic for $t > 0$, we have $\overline{K_T(\mathbf{F})} = \overline{\bigcup_{0<t} K_t(\mathbf{F})}$, with T arbitrary, $0<T<\infty$, so that approximate controllability in $(0,T)$ and in finite time are the same concept for $\mathbf{F}$. We shall therefore often omit the time length specification for $\mathbf{F}$ and talk only about its approximate controllability.

We now collect below a series of elementary but important results to be used in the sequel.

Claim (i). It follows easily, via (2.4) with $x_0 = 0$ and a direct application of a standard consequence (H1, p.31) of the Hahn-Banach theorem, that $\mathbf{F}$ is approximately controllable in $(0,T)$ if and only if

$$\forall x^* \varepsilon X^*,\ x^*(S(t)BU) \equiv 0,\ 0 \leq t \leq T \Rightarrow x^* = 0 \qquad (2.5)$$

For $\mathbf{F}_m$ the above characterization specializes to

$$\forall x^* \varepsilon X^*,\ x^*(S(t)b_i) \equiv 0,\ 0 \leq t \leq T,\ i = 1,\dots,m \Rightarrow x^*=0 \qquad (2.5')$$

For an equivalent formulation see (F3).

Remark 2.1. It is plain from the characterization (2.5) that $\mathbf{F}(A,B)$ is approximately controllable if and only if so is $\mathbf{F}(A,S(\bar{t})B)$ for any $\bar{t} \geq 0$

Claim (ii). Similarly, via (2.1) and (2.2) with $x_1,= x_0 = 0$, $\mathbf{S}$ is approximately controllable in $(0,T)$ (resp. in finite time) if and only if

$$\begin{aligned}&\forall x_1^*, x_2^* \varepsilon X^* \\ &\int_0^t x_1^*(C(\tau)BU)d\tau + x_2^*(C(t)BU \equiv 0,\ 0\leq t\leq T,\ (\text{resp. } t\geq 0) \Rightarrow x_1^* = x_2^* = 0\end{aligned} \qquad (2.6)$$

For $\mathbf{S}_m$ the above characterization specializes to

$$\begin{aligned}&\forall x_1^*, x_2^* \varepsilon X^* \\ &\int_0^t x_1^*(C(\tau)b_i)d\tau + x_2^*(C(t)b_i) \equiv 0,\ i=1,\dots,m;\ 0\leq t\leq T;\ (\text{resp. } t\geq 0) \Rightarrow x_1^* = x_2^* = 0\end{aligned} \qquad (2.6')$$

For an equivalent formulation see (F1).

LEMMA 2.1. Under assumption H1, the following holds for any $y \varepsilon X$:

$\mathrm{Cl}\{C(t)y,\ 0\leq t<\infty\} = \mathrm{Cl}\{S(t)y,\ 0 \leq t<\infty\} = \mathrm{Cl}\{S(t)y,\ 0\leq t\leq T\}$, T arbitrary, $0<T<\infty$

where here and hereafter Cl denotes 'closure'.

Remark 2.2. Since $C(t)$ is an even function of t (from the D'Alambert functional identity (F2;K4)), the set on the left hand side does not change if t runs over the entire real line, even if the sign Cl is suppressed.

Proof. By the Hahn-Banach theorem, all we have to show to justify the first equality is that if $x^* \varepsilon X^*$ and $x^*(S(t)y) \equiv 0$, $t\geq 0$, then also $x^*(C(t)y) \equiv 0$, $t\geq 0$; and conversely. Both implications follow from (2.3), the first one, by a change of variable, as in the uniqueness of the Laplace transform (D1, p.626). The second equality is due to the analyticity of $S(t)$. Q.E.D.

For later reference we label another condition:

$$\forall x^* \varepsilon X^*,\ x^*(C(t)BU) \equiv 0,\ 0\leq t<\infty, \Rightarrow x^* = 0. \qquad (2.7)$$

Now, let $\mathbf{F}$ fail to be approximately controllable in $(0,T)$, hence in finite time. Then, by claim (i) above, $\overline{x}^*(S(t)BU) \equiv 0$, $t\geq 0$, for some non zero $\overline{x}^* \varepsilon X^*$. By Lemma 2.1, $\overline{x}^*(C(t)BU) \equiv 0$, $t\geq 0$ and so

$\int_0^t \overline{x}^*(C(\tau)BU)d\tau \equiv 0,\ t\geq 0$ and the left hand side of the implication (2.6) is violated with $x_1^* = x_2^* = \overline{x}^* \neq 0$. We have thus proved

THEOREM 2.2. Let A satisfy H1. a) If $\mathbf{S}$ is approximately controllable in finite time, then $\mathbf{F}$ is approximately controllable in any $(0,T)$, $0<T<\infty$. In symbols: $\overline{\bigcup_{0\leq t<\infty} K_t(\mathbf{S})} = X \times X \Rightarrow \overline{K}_T(\mathbf{F}) = X$. b) The reverse implication also holds, if condition (2.7) implies condition (2.6).

Remark 2.3. Eq. (2.6) always implies Eq. (2.7) for any T, while the converse is not always true. However, since $C(t)$ is even and so $\int_0^t C(\tau)d\tau$ is odd on $(-\infty,\infty)$, one also has that Eq. (2.7) implies Eq. (2.6), provided (2.6) for $t\geq 0$ can be extended to hold also for all negative values of t. This is the case e.g., when A is bounded on X.

Before stating the next result, we define $D_\infty(A) = \bigcap_{n=1}^\infty D(A^n)$ and recall that, for an operator A generating a C_0-semigroup, hence a fortiori for A satisfying H1, $D_\infty(A)$ is a subspace and is still dense in X (B1,p. 12). Also, let $X_0 = \bigcup_{0<t} S(t)X$, so that $b_i \in X_0$ is case $b_i = S(\tilde{t}_i)\tilde{b}_i$, for some $\tilde{b}_i \in X$ and $\tilde{t}_i>0$. X_0 is a subspace dense in X (H1, p. 208). Notice that $X_0 \subset D_\infty(A)$ for any semigroup satisfying $S(t)X \subset D(A)$ for $t>0$, i.e. for any differentiable semigroup, hence a fortiori for the analytic semigroup generated by A satisfying H1. (In fact: $S(t)X \subset D(A)$ implies $AS(t)X = S(t/2)AS(t/2)X \subset D(A)$ and hence $S(t)X \subset D(A^2)$ etc. by induction.) That condition H2 is not necessary for the approximate controllability of $\mathbf{F}(A,B)$ to imply the same property for $\mathbf{S}(A,B)$ will be proved in Corollary 2.4 below. To this end, we shall focus our attention on smooth vectors of X contained in $D_\infty(A)$. The next result is in the spirit of the extension, as given by the author in (T2, Sect.2), of the classical rank condition for controllability of finite dimensional systems, of which it is a generalization.

THEOREM 2.3. Let A satisfy H1.

a) With $BU \subset D_\infty(A)$ (resp. with $b_i \in D_\infty(A), i=1,..,m$), the condition

$$\overline{\mathrm{sp}}\{A^n BU\}_{n=0}^\infty = X \qquad (\text{resp. } \overline{\mathrm{sp}}\{A^n b_i, i=1,..,m\}_{n=0}^\infty = X)$$

is sufficient for approximate controllability on any $(0,T)$, $0<T<\infty$, of $\mathbf{S}(A,B)$ (resp. $\mathbf{S}(A,(b_1,..,b_m))$), hence of $\mathbf{F}(A,B)$ (resp. $\mathbf{F}_m(A,(b_1,..,b_m))$). Such condition is however not even necessary for approximate controllability of $\mathbf{F}_m(A,(b_1,..,b_m))$.

b) With $BU \subset X_0$ (resp. with $b_i \in X_0, i=1,..,m$), the condition

$$\overline{\mathrm{sp}}\{A^n BU\}_{n=0}^\infty = X \qquad (\text{resp. } \overline{\mathrm{sp}}\{A^n b_i, i=1,..,m\}_{n=0}^\infty = X).$$

is necessary and sufficient for approximate controllability on any $(0,T)$, $0<T<\infty$, of both $\mathbf{S}(A,B)$ (resp. $\mathbf{S}_m(A,(b_1,..,b_m))$) as well as $\mathbf{F}(A,B)$ (resp. $\mathbf{F}_m(A,(b_1,..,b_m))$).

COROLLARY 2.4. Let A satisfy H1. With $BU \subset X_0$ (resp. with $b_i \in X_0$, $i=1,..,m$), approximate controllability on any $(0,T)$, $0<T<\infty$, of $\mathbf{F}(A,B)$ (resp. $\mathbf{F}_m(A,(b_1,..,b_m))$) implies the same property for $\mathbf{S}(A,B)$ (resp.

$\mathbf{S}_m(A,(b_1,..,b_m)))$.

Remark 2.4. If the operator B of the original approximately controllable system $\mathbf{F}(A,B)$ does not satisfy $BU \subset X_0$, reference to Remark 2.1 is useful. Similarly for F_m.

Proof of Theorem 2.3. Part a). That the stated condition is not necessary for approximate controllability of $\mathbf{F}_m(A,(b_1,..,b_m))$ was already pointed out by the author in (T2, Remark 2.4) by means of an example with $m=2$ (Example 2.1 of the present paper in fact). As for the sufficiency in part a) for $\mathbf{F}_m$, let, by contradiction, there exist $\vec{x}_1^*$, $\vec{x}_2^* \in X^*$, not both zero, such that (see Claim (ii))

$$\int_0^t \vec{x}_1^*(C(\tau)b_i)d\tau + \vec{x}_2^*(C(t)b_i) \equiv 0, \quad 0\le t\le T,\ i=1,..,m \tag{2.8}$$

Recall now (e.g. K2) that if $y \in D_\infty(A)$ then

$$\frac{d^{2n}C(t)y}{dt^{2n}} = A^nC(t)y = C(t)A^ny \qquad n=1,2,3,..\ -\infty<t<\infty \tag{2.9}$$

Also

$$C(0)x = x,\ x\in X \quad \text{and} \quad \left.\frac{dC(t)x}{dt}\right|_{t=0} = 0,\ x \in D(A) \tag{2.10}$$

Setting $t=0$ in (2.8) yields $\vec{x}_2^*(b_i) = 0$. Differentiate successively (2.8) using (2.9), and set $t=0$ at each step using (2.10) to show, by induction, that

$$\vec{x}_1^*(A^nb_i) = 0 \quad \text{and} \quad \vec{x}_2^*(A^nb_i) = 0, \quad n=0,1,... \tag{2.11}$$

But in view of the Hahn-Banach theorem, this contradicts the hypothesis, since $\vec{x}_1^*$ and $\vec{x}_2^*$ are not both zero. Similarly for $\mathbf{F}$. Since $X_0 \subset D_\infty(A)$ as previously observed, it remains to show necessity for part b). To this end, let $b_i \in X_0$, i.e. $b_i = S(\tilde{t}_i)\tilde{b}_i$, for some $\tilde{t}_i > 0$ and $\tilde{b}_i \in X$ and let

$$\vec{x}^*(A^nb_i) \equiv \vec{x}^*(A^nS(\tilde{t}_i)\tilde{b}_i) \equiv 0,\ i=1,..,m;\ n=0,1,...$$

for some nonzero $\vec{x}^* \in X^*$. Then, by the analyticity of $S(t)$ which implies (B1, pp. 15-16)

$$\frac{d^nS(t)}{dt^n} = A^nS(t), \quad t>0 \tag{2.12}$$

it follows that $\vec{x}^*(S(t)\tilde{b}_i) \equiv 0$ as well as $\vec{x}^*(S(t)b_i) \equiv 0$, $0\le t<\infty$. By Lemma 2.1, one also has $\vec{x}^*(C(t)b_i) \equiv 0$, $-\infty<t<\infty$. Applying the previous Claims (i) and (ii) yields easily a contradiction with the assumptions, since $\vec{x}^*$ is nonzero. Q.E.D.

Remark 2.5. Recall Fattorini's example from section 1. Writing $A_r = \Delta + rI$, we have $S_{A_r}(t) = e^{rt}S_\Delta(t)$ with

$$(S_\Delta(t)f)(\xi) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-\frac{(\xi-\sigma)^2}{4t}} f(\sigma)d\sigma, \quad t>0 \quad -\infty<\xi<\infty \tag{2.13}$$

(B1, p. 2). Then $b_1(\xi) = e^{-|\xi|}$ as well as $b_2(\xi) = b_1(\xi+1)$ do not belong to the subspace $X_0 = \bigcup_{0<t} S_{A_r}(t)X$ in agreement with our Theorem 2.3. This is so, since $b_1(\xi)$ has a discontinuity in its first derivative at the origin, while this does not occur for the function in ξ defined by the right hand side of (2.13), for any $f(\cdot) \varepsilon L_2(-\infty,\infty)$.

Remark 2.6. When $BU \subset D_\infty(A)$, the following inclusion is contained in the proof of Thm. 2.3a)

$$\overline{sp}\{A^nBU\}_{n=0}^{\infty} \subset Cl\{S(t)BU,\ 0\leq t<\infty\} \tag{2.14}$$

The next two examples with $m=2$ illustrate the case when the inclusion is proper, with the set on the r.h.s. being the entire space:

$$\overline{sp}\{A^nb_i, i=1,2\}_{m=0}^{\infty} \subsetneqq Cl\{S(t)b_i, i=1,2;\ 0\leq t<\infty\} = X$$

Example 2.1. Let $X = L_2(-\infty,\infty)$, $Af = d^2f/d\xi^2$ (in the sense of distributions) with $D(A) = \{f\colon f \text{ and } f'' \varepsilon L_2(-\infty,\infty)\}$. Take $b_1(\xi)$ to be a C^∞ function with compact support. Say: $b_1(\xi) = \exp((\xi^2-1))^{-1}$, $-1 < \xi < 1$, and $b_1(\xi) = 0$ for $|\xi| \geq 1$. Define $b_2(\xi) = b_1(\xi+h)$, $h \neq 0$. Then $b_1(\cdot)$ and $b_2(\cdot)$ belong to $D_\infty(A)$ and, moreover, they vanish identically together with all their derivatives outside $(-1,1+h)$ for $h > 0$. ($(-1+h,1)$ for $h > 0$). Hence in this case we have $\overline{sp}\{A^nb_i,\ i = 1,2\}_{n=0}^{\infty} \subsetneqq X$ and there is a non zero $\overline{x}^* \varepsilon X^*$ such that $\overline{x}^*(A^nb_i) \equiv 0$, $i = 1,2$; $n = 0,1,\ldots$ Yet the identity $\overline{x}^*(S(t)b_i) \equiv 0$, $0\leq t<\infty$, $i = 1,2$ is false in view of the previous Claim (i) since, as shown in (F2, Ex.1; T3, Ex.3.1) the present system $\mathbf{F}_2(A,(b_1,b_2))$ is indeed approximately controllable. Notice that $b_i \not\in X_0$, in agreement with Thm. 2.3b).

Example 2.2. With X and A as in the previous Example 2.1, let $b_1(\xi) = e^{-|\xi|}$ and $b_2(\xi) = b_1(\xi+1)$, as in Fattorini's example. Then $b_1(\cdot)$ and $b_2(\cdot)$ belong to $D_\infty(A)$ since the derivatives are taken in the sense of distribution (A1,p. 21). Also, $\overline{sp}\{A^nb_i,\ i = 1,2\}_{n=0}^{\infty} \subsetneqq X$ since otherwise, by theorem 2.3a), we would have that the corresponding system $\mathbf{S}_2(A,(b_1,b_2))$ is approximately controllable on any $(0,T)$ which we know from (F1) to be false. So there is a nonzero $\overline{x}^* \varepsilon X^*$, such that $\overline{x}^*(A^nb_i) \equiv 0$, $i=1,2$; $n=0,1,\ldots$ Yet the identity $\overline{x}^*(S(t)b_i) \equiv 0$, $0\leq t<\infty$ is false, since as noted in (F1) the corresponding system $\mathbf{F}_2(A,(b_1,b_2))$ is indeed approximately controllable.

Remark 2.7. Theorem 2.3b) on the other hand shows that, if in particular $BU \subset X_0$ (resp. $b_i \varepsilon X_0$, $i=1,\ldots,m$), then

$$\overline{sp}\{A^nBU\}_{n=0}^{\infty} = Cl\{S(t)BU,\ 0\leq t<\infty\} \tag{2.15}$$

(resp.

$$\overline{sp}\{A^nb_i,\ i=1,\ldots,m\}_{n=0}^{\infty} = Cl\{S(t)b_i,\ i=1,..,m;\ 0\leq t<\infty\} \tag{2.15'}$$

COROLLARY 2.5. Let A be bounded on X. Then $\mathbf{S}(A,B)$ is approximately controllable $(0,T)$ if and only if so is $\mathbf{F}(A,B)$ and this happens just in case

$$\overline{sp}\{A^n BU\}_{n=0}^{\infty} = X \tag{2.16}$$

Remark 2.8. When A is bounded on X (in which case assumption H2 is certainly satisfied), then $S(t) = e^{At} = \sum_{n=0}^{\infty} A^n t^n/n!$ and $C(t) = \sum_{n=0}^{\infty} A^n t^{2n}/(2n)!$, $-\infty<t<\infty$ and the above argument leading to Corollary 2.5 simplifies, by making direct use of the explicit power series expansions for $S(t)$ and $C(t)$. Also, the characterization (2.16) was already given in (T1, Thm. 3.1.1) for **F**. We think therefore, that it is instructive to insert another proof, which deduces the statement for **S** in Corollary 2.5 from the correspondent known (T1) statement for **F**. Using a standard device **S** can be written as first order system on $X \times X$

$$\dot{y} = \tilde{A}y+\tilde{B}v, \quad y \in Y = X \times X, \quad v \in V = U\times U \text{ with} \tag{2.17}$$

$$\tilde{A} = \begin{vmatrix} 0 & I \\ A & 0 \end{vmatrix}, \qquad \tilde{B} = \begin{vmatrix} 0 \\ B \end{vmatrix}$$

bounded operators on Y and $V \to Y$, respectively. Compute the powers $\tilde{A}^n\tilde{B}V$, $n=0,1,\dots$ to get

$$\begin{vmatrix} 0 \\ BU \end{vmatrix} ; \begin{vmatrix} BU \\ 0 \end{vmatrix} ; \begin{vmatrix} 0 \\ ABU \end{vmatrix} ; \begin{vmatrix} ABU \\ 0 \end{vmatrix} ; \begin{vmatrix} 0 \\ A^2BU \end{vmatrix} ; \begin{vmatrix} A^2BU \\ 0 \end{vmatrix} , \text{ etc.} \tag{2.18}$$

From the quoted result for first order systems, approximate controllability of (2.17) - i.e. of **S** - is equivalent to

$$\overline{\text{sp}}\{\tilde{A}^n\tilde{B}V\}_{n=0}^{\infty} = X \times X$$

which, in view of (2.18), is in turn equivalent to (2.16)

The next corollary is a comprehensive statement containing previous known results for first order systems ((F3, Prop. 2.3) and (T1, Thm. 3.1.1)) together with Theorems 2.1 and 2.3 above as well as (F1, Thm. 2.3) for second order systems.

COROLLARY 2.6. Let A satisfy H1 and consider the following four systems

$$\mathbf{F}\ (A,B) : \dot{x} = Ax + Bu \quad ; \quad \mathbf{F}\ (R(\lambda_0,A),B) : \dot{x} = R(\lambda_0,A)x + Bu$$

$$\mathbf{S}\ (A,B) : \ddot{x} = Ax + Bu \quad ; \quad \mathbf{S}\ (R(\lambda_0,A),B) : \ddot{x} = R(\lambda_0,A)x + Bu$$

where λ_0 is fixed but otherwise arbitrary point in $\rho_0(A)$ (see§1) with $\text{Re}\lambda_0 > \omega_0^2$. Then

a) either the systems $\mathbf{F}(A,B)$, $\mathbf{F}(R(\lambda_0,A),B)$ and $\mathbf{S}(R(\lambda_0,A),B)$ are all simultaneously approximately controllable on any $(0,T)$, $0 < T<\infty$, or none of them is; the first alternative occurs if and only if

$$\overline{\text{sp}}\{R^n(\lambda_0,A)BU\}_{n=0}^{\infty} = X \tag{2.19}$$

b) Eq. (2.19) is a necessary condition for approximate controllability in finite time of $\mathbf{S}(A,B)$ and it is also sufficient in the following two cases

(i) either A satisfies Fattorini's assumption H2

(ii) or else (with no assumption on A except H1) the range of B is contained in $X_0 = \bigcup_{0<t} S(t)X$.

3. Comparison with Tsujioka's work

After the first draft of the present paper was completed (during a one-month stay in August 1975, at the Mathematics Research Center, University of Wisconsin, Madison, sponsored by the United States Army, under contract number DAAG29-75-C0024, whose financial support is gratefully acknowledged), H. O. Fattorini kindly brought to our attention reference (T1) by Tsujioka, which tackles the same problem with a slightly different terminology from ours. Therefore, a comparison between Tsujioka's work and the present paper is in order.

1. First of all, there is a difference in scope and generality, in the sense that Tsujioka's paper is definitely more restrictive: in fact it investigates only the special case when the space X is Hilbert and A is a selfadjoint operator semibounded above. Our present paper instead treats the most general operator A for which the problem under consideration makes sense on an arbitrary Banach space (i.e. an operator A satisfying H1).

2. Second, there is a difference in the approach or methodology being employed. In fact, by making use of his more restrictive assumptions, Tsujioka rewrites the second order system as a first order system in the product space. $H_{\frac{1}{2}} \times X$, by means of the standard device of introducing the velocity vector as a new coordinate. $H_{\frac{1}{2}}$ is the domain $D(A_\alpha^{\frac{1}{2}})$ endowed with norm $\|x\|_{H_{\frac{1}{2}}} = \|A_\alpha^{\frac{1}{2}}x\|_X$, where $A_\alpha^{\frac{1}{2}}$ is the positive square root of the positive selfadjoint operator $A_\alpha = -A+\alpha$, for some suitable real constant α for which $((-A+\alpha)x,x)_X \geq \delta \|x\|_X^2$, $\delta > 0$. $H_{\frac{1}{2}}$ becomes a Hilbert space with inner product $(x,y)_{H_{\frac{1}{2}}} = (A_\alpha^{\frac{1}{2}}x, A_\alpha^{\frac{1}{2}}y)_X$ for $x,y \in D(A_\alpha^{\frac{1}{2}})$. (Appendix 1) On the product space $H_{\frac{1}{2}} \times X$, the corresponding first order system defines a C_0-group, generated by the operator $\begin{vmatrix} 0 & I \\ A & 0 \end{vmatrix}$ with domain $D(A) \times D(A_\alpha^{\frac{1}{2}})$. The general problem of rewriting the second order equation as a first order equation is considered e.g. in (F2,G1,K3) and requires picking up correctly the space X_1 and its norm. It is not always true that the space X_1 can be taken to be equal to X, without losing the uniform well-posedness of the new first order equation^{3} (F2). (For instance, it is well known that, in the case of the wave equation $A = \Delta$ on R^n, the space on which the operator $\begin{vmatrix} 0 & I \\ \Delta & 0 \end{vmatrix}$ generates a C_0-group is $H^1(R^n) \times L_2(R^n)$). For this reason we purposely take the more general approach, which appears to us more natural, of leaving the second order equation as such, without rewriting it as first order equation and study the controllability problem in the product space $X \times X$ for the pair made of the position vector and the velocity vector. Our analysis therefore is carried out using the theory of strongly continuous abstract cosine functions. In the literature, e.g. in (G1) one also finds that the space X_1 is taken to be $D(A_\alpha^{\frac{1}{2}})$ equipped with the graph norm. In Appendix 2, we show the equivalence between the graph norm and the $H_{\frac{1}{2}}$-norm on $D(A_\alpha^{\frac{1}{2}})$.

In Tsujioka's approach, reducing the second order system to a first order system on $H_{\frac{1}{2}} \times X$, the controllability requirement is then naturally to demand that the totality of pairs of position and velocity vectors, corresponding to all admissible controllers over all nonnegative times, be dense in $H_{\frac{1}{2}} \times X$. In our approach, we demand instead that such totality be dense in $X \times X$. It is then natural to ask what is the relationship between these two denseness requirements. We prove in Appendix 3 that they are, in fact, topologically equivalent.

3. Third, the main result of (T1) i.e. its Thm. 2, which Tsujioka largely credits to the referee, is indeed related to our Corollary 2.4. However, even under the more restrictive assumption of selfadjointness (Thm 2, T1) does not dispel the doubt - as we do here in a more general context - that Fattorini's assumption H2 is not necessary for the implications (in our terminology); "$\mathbf{F}(A,B)$ approximately controllable (on any $(0,T)$) $\Rightarrow$ ($\mathbf{S}(A,B)$ approximately controllable in finite time." In fact, assuming $\mathbf{F}(A,B)$ approximately controllable, Tsujioka seeks another operator B', different from B, such that $\mathbf{S}(A,B')$ is also approximately controllable on any $(0,T)$. Thm. 2 of (T1) asserts that such operator B' can be taken to be $B' = CB$, where C is a bounded operator on X, satisfying the assumptions a),b),c),d) of (Thm. 2, T1). In particular C may be taken to be $C = S(\varepsilon)$, $\varepsilon > 0$; it is here that one sees that Thm. 2 in (T1) is close to our Corollary 2.4 (see also our Remark 2.1), at least in the selfadjoint case.

4. The important observation that the subspace $X_0 = \underset{0<t}{\cup} S(t)X$ is dense in X is also missing in (T1). This is crucial in appreciating our Corollary 2.4: if one is allowed to modify, say, the original vectors b_i of the models $\mathbf{F}(A,(b_1,\ldots,b_m))$ and $\mathbf{S}(A,(b_1,\ldots,b_m))$, all that is needed is to perturb them by an arbitrarily small perturbation as to achieve that the new b_i's be in X_0.

5. Finally, (T1) does not bring any insight into the problem under investigation - as we do here - from the viewpoint of generalizing to infinite dimensional systems the familiar finite dimensional rank condition for controllability.

FOOTNOTES

1. Eq. (2.3) is the generalization to Banach spaces and unbounded operators of the scalar formula

$$e^{at} = \frac{1}{\sqrt{\pi t}} \int_0^\infty e^{-\frac{\tau^2}{4t}} \cos(-a)^{1/2}\tau \; d\tau, \; t>0$$

valid for an arbitrary complex constant a.

2. It is plain from (2.1) and (2.4) that the closure of the set of attainability $K_t(\cdot)$ does not change if, instead of taking $L_1((0,t),U)$-control functions, one restricts to any other linear class of control functions which are dense in $L_1((0,t),U)$.

3. However X_1 can be taken equal to X, when A is bounded as in Remark 2.8.

4. Appendices

Appendix 1. Completeness of $H_{\frac12}$.

Let $\{x_n\}$, $x_n \varepsilon D(A_\alpha^{\frac12})$ be a Cauchy sequence with respect to the norm $\| \ \|_{H_{\frac12}}$. Equivalently, $\{A_\alpha^{\frac12} x_n\}$ is a Cauchy sequence with respect to the norm $\| \ \|_X$, and so $\|A_\alpha^{\frac12} x_n - y\|_X \to 0$ as $n \to \infty$ for some $y \varepsilon X$. Then $(A_\alpha^{\frac12})^{-1} y \varepsilon D(A_\alpha^{\frac12})$ and $\|x_n - (A_\alpha^{\frac12})^{-1} y\|_{H_{\frac12}} \to 0$ as $n \to \infty$ which proves completeness of $H_{\frac12}$. Q.E.D.

Appendix 2. On $D(A_\alpha^{\frac12})$ the norms $\|x\|_{H_{\frac12}}$ and the graph norm $|||x|||_G \overset{d}{=} \|A^{\frac12} x\|_X + \|x\|_X$ are equivalent.

In fact one has immediately $|||x|||_G \geq \|x\|_{H_{\frac12}}$. Also, since $(A_\alpha^{\frac12})^{-1}$ exists as a bounded operator on X, one has $\|x\| = \|(A_\alpha^{\frac12})^{-1} A^{\frac12} x\|_X \leq \|(A_\alpha^{\frac12})^{-1}\| \, \|x\|_{H_{\frac12}}$ and hence $|||x|||_G \leq (1 + \|(A^{\frac12})^{-1}\|)\|x\|_{H_{\frac12}}$. Q. E. D.

Appendix 3. Denseness on X in the $\| \ \|_X$-norm is equivalent to denseness on $H_{\frac12}$ in the $\| \ \|_{H_{\frac12}}$-norm.

Assume denseness on X in the $\| \ \|_X$-norm and let x be an arbitrary vector in $D(A_\alpha^{\frac12})$. Consider $A_\alpha^{\frac12} x$ and let $\varepsilon > 0$ be given. Then there is some y in X such that

$$\|x - (A_\alpha^{\frac12})^{-1} y\|_{H_{\frac12}} = \|A_\alpha^{\frac12} x - y\|_X < \varepsilon$$

and $(A_\alpha^{\frac12})^{-1} y \varepsilon D(A_\alpha^{\frac12})$ is within ε from x. Conversely, assume denseness on $H_{\frac12}$ in the $\| \ \|_{H_{\frac12}}$-norm. Let y be an arbitrary vector in X and let $\varepsilon > 0$ be given. There is x in $D(A_\alpha^{\frac12})$ such that $\|y - x\|_X < \varepsilon/2$. But then, for such x, there is z in $D(A_\alpha^{\frac12})$ such that

$$\|A_\alpha^{\frac12}(x - z)\|_X = \|x - z\|_{H_{\frac12}} < \frac{\varepsilon}{2\|(A_\alpha^{\frac12})^{-1}\|}$$

Therefore

$$\|x - z\|_X = \|(A_\alpha^{\frac12})^{-1} A_\alpha^{\frac12}(x - z)\|_X \leq \|(A_\alpha^{\frac12})^{-1}\| \, \|A_\alpha^{\frac12}(x - z)\| = \frac{\varepsilon}{2}$$

and $\|y - z\|_X < \varepsilon$ Q.E.D.

Remark A.1. All the above proofs do not exploit the selfadjointness of the operator A, but only the fact that a suitable translation of $-A$ admits a positive square root.

REFERENCES

(A1) R. A. Adams, Sobolev spaces, Academic Press 1975.

(B1) P. L. Butzer and H. Berens, Semigroups of Operators and Approximations, Springer-Verlag, Berlin 1967.

(D1) N. Dunford and J. T. Schwartz, Linear Operators, Parts 1 and 2, Interscience New York 1959 and 19 .

(F1) H. O. Fattorini, Controllability of higher order linear systems, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustad, eds., Academic Press, New York 1967.

(F2) H. O. Fattorini, Ordinary differential equations in linear topological spaces, I and II, J. Differ. Equat. 5(1968), p.72-105 and 6(1969), pp. 50-70.

(F3) H. O. Fattorini, On complete controllability of linear systems, J. Differ. Equat. 3(1967), pp. 391-402.

(F4) H. O. Fattorini, Some remarks on complete controllability SIAM J. Control. 4(1966) pp. 686-694.

(G1) J. Goldstein, On a connection between first and second order differential equations in Banach space, J. of Math. Anal. and Applic. 30 (1970) pp.246-251.

(H1) E. Hille and R. S. Phillips, Functional Analysis and semigroups, American Mathematical Society, Providence, R.I. 1958.

(K1) T. Kato, Perturbation theory of linear operators, Springer-Verlag, Berlin 1966.

(K2) J. Kisinsky, On cosine operator functions and one parameter groups of operators, Studia Mathematica, T. XLIV (1972).

(K3) J. Kisinsky, On the connection between cosine operator functions and one parameter semigroups and groups of operators, University of Warsaw (Poland), Institute of Mathematics Report, 1972.

(K4) J. Kisinsky, On operator-valued solutions of D'Alambert's functional equation, II, Studia Mathematica T. XLII. (1972), pp.43-66.

(T1) K. Tsujioka, Remars on controllability of second order evolution equations in Hilbert spaces, SIAM J. Control, Vol. 8, No. 1, Febr. 1970, pp. 90-99.

(T2) R. Triggiani, Controllability and Observability in Banach space with bounded operators, SIAM J. Control 13(1975) pp. 462-491.

(T3) R. Triggiani, Extensions of rank conditions for controllability and observability to Banach spaces and unbounded operators, SIAM J. Control, 14(1976), pp. 313-338.

OPERATOR THEORY AND BROADBAND MATCHING

J. WILLIAM HELTON
University of California, San Diego, La Jolla, California
Department of Mathematics

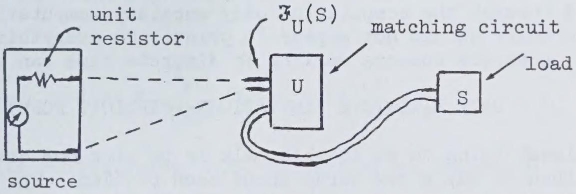
If one connects a power source to a frequency dependent load through a lossless coupling circuit, the amount of energy transferred to the load naturally depends on the characteristics (frequency response function) of the coupling circuit. The problem is to find which coupling circuit transfers the most power to the load. Presently there is a reasonable theory (with a few gaps) for lumped one ports.

The present work uses operator theory techniques to solve the problem for strictly passive n-port loads. We also indicate how to handle passive loads and fill in what the author considers to be gaps in the one port theory. Most techniques carry over to infinite dimensional systems easily. The results actually described in this talk are a subset of those given in [H], and a seriously interested reader should consult it.

- - - - - -

It has been known for several years that the 'canonical models' branch of operator theory (a close relative of Lax-Phillips scattering theory) is intimately connected with the engineering study of 'systems', c.f., IEEE Proceedings, Jan. 1976. The implication is not only that mathematical results have engineering implications in a direct way (and conversely), but that the operator theory techniques have considerable power for attacking systems and network problems. Today I will describe such an approach to the classical broadband matching problem.

The problem is this: one has a load system (with 'scattering matrix' S) and wishes to drive it with a power source having unit internal impedance. The source can be connected to the load system through an intermediate lossless (matching) system (with 'scattering matrix' U)



and the problem is to find the matching circuit U which delivers the most power to S. A variation is to find U which delivers a prescribed amount of power to S at each fixed frequency. A related problem is the problem of 'compatible impedances,' which amounts to fixing S and trying to determine all circuits which can be made from it by cascading with a lossless U. The most famous special case has S = 0 (it is a unit resistor) and is called Darlington synthesis. One can work with one ports, n-ports, or even ∞-ports. Theoretical engineering work here begins with Bode and continues with Fano, Youla, and many others. The approach in my study is closest to Youla's and the relationships will be described in more detail later.

THE MATHEMATICAL PROBLEM

Let $H^\infty(\mathbb{C}^n)$ denote the uniformly bounded analytic functions in the disk with matrix values and let $\mathcal{B}$ denote the open unit ball in that space, namely, $\mathcal{B} = \{f \in H^\infty(\mathbb{C}^n): \|f(z)\| \leq \alpha < 1$ for some α and all $|z| \leq 1\}$. The closed unit ball is denoted $\overline{\mathcal{B}}$. Define a *cascade transformation* to be a map $\mathfrak{F}_U: \mathcal{B} \to \mathcal{B}$ which is defined for each $S \in \mathcal{B}$ by

$$\mathfrak{F}_U(S)(z) = a(z) + c(z)S(z)\,(I-d(z)S(z))^{-1}b(z) \, ,$$

where the function

$$U(z) = \begin{pmatrix} a(z) & b(z) \\ c(z) & d(z) \end{pmatrix}$$

satisfies $d(z) \not\equiv 0$ and is an *inner* function, namely it is in $\overline{\mathcal{B}}(\mathbb{C}^{2n})$ and has unitary boundary values. The fact that a cascade transformation maps $\mathcal{B}$ into itself (it also maps $\overline{B}$ into itself) is easily verified (c.f., [H-Z]) as is the fact that the set $\mathcal{G}$ of all cascade transformations is a semigroup. The subset $\{\mathfrak{F}(S): \mathfrak{F} \in \mathcal{G}\}$ of $\mathcal{B}$ is called the *orbit of* S *under* $\mathcal{G}$. Traditionally, in studying transformation groups mathematicians devote considerable attention to their orbit structure and that just happens to be what the engineering problems amount to. In particular this article concerns two types of questions:

I. What is the orbit structure of $\mathcal{G}$ on $\overline{\mathcal{B}}$? That is, given S describe all Σ in its orbit (problem of compatible impedances).

II. Given S, which point Σ in its orbit is closest to 0? What is the distance? Which $\mathfrak{F}_U$ takes S to Σ? (Broadband matching problem)

Let me apologize for the 'unengineeringness' of the conventions. Here all our frequency response functions are analytic on the unit disk Δ rather than the R.H.P. Of course, one can Cayley transform Δ onto the R.H.P. and thereby convert results on Δ to results on R.H.P. As of this writing I have worked through the computation only once;any computation this author hasn't done twice should not appear in print. So everything herein will be set on Δ, possibly someone working in discrete time can use this directly.

GAIN BANDWIDTH LIMITATIONS (EXPLICIT FORM)

The wisest thing to do in this talk is to pick one question and answer it rather than to say a few words about each problem. The one we emphasize is: What is the distance from O_S to 0? First of all let us discuss what is meant by distance. Suppose μ is a function $0 \leq \mu(e^{i\theta}) \leq 1$ which serves to weight each frequency. A weight close to 0 will mean the frequency is important and should receive lots of power; a weight near 1 means the opposite. Let $L^\infty(\mathbb{C}^n)$ denote the $n \times n$ matrix valued functions on $T \overset{\Delta}{=} \{z: |z|=1\}$ which are uniformly bounded and define a norm on ρ_μ on $L^\infty(\mathbb{C}^n)$ by

$$\rho_\mu(f) = \sup_{e^{i\theta}} \frac{1}{\mu(e^{i\theta})} \|f(e^{i\theta})\|_{\mathbb{C}^n} \, .$$

Note that $\rho_\mu(\Sigma) \leq \beta$ says

$$\|\Sigma(e^{i\theta})\| \leq \beta\ \mu(e^{i\theta})\ ,$$

for almost all θ. So computing the ρ_μ distance of 0_S to zero amounts to finding Σ in 0_S which satisfy this type of inequality. We always take μ rational.

The plan for this presentation is to present special results first in an explicit form. After that describe some of the ingredients of the proof and then give more general results in a more elegant form. We say that M is a left (resp. right) spectral factor of $P \in L^\infty(\mathbb{C}^n)$ if M is in $H^\infty(\mathbb{C}^n)$, and

$$M(E^{i\theta})M(e^{i\theta})^* \qquad (\text{resp. } M(e^{i\theta})^*M(e^{i\theta}))\ .$$

Suppose we are given a bounded real strictly passive $S(z) = \frac{g(z)}{q(z)}$ where g,q are relatively prime polynomials. Also suppose we are given a rational weight function μ. Define

$$S_c(z) = z^N \frac{g\left(\frac{1}{z}\right)}{q(z)} \quad \text{where } N \text{ is the order of } g\ ,$$

$$\alpha = \text{the spectral factor of } \mu^2 \text{ with no zeroes in } \Delta\ .$$

We also let T_S denote the *transmission* zeroes of S, namely, it is the set of poles z_k in Δ of $S\left(\frac{1}{z}\right)\left(1 - S\left(\frac{1}{z}\right)S(z)\right)^{-1}$ together with their multiplicities r_k; written $\{[z_k,r_k]\}$.

Finally we define the *Pick matrix* $\Lambda_{T_S}(\psi)$ for the 'multi' set T_S and rational function ψ by the following recipe (we always suppose ψ is analytic at each point of T_S):

For $\lambda \in T_S$ define $\delta_\lambda(z) \triangleq \frac{z-\lambda}{1-z\bar{\lambda}}$ and expand

$$\psi(z) = \sum_{n=0}^{\infty} b_n(\lambda)(\delta_\lambda(z))^n\ .$$

For $\lambda,\mu \in T_S$ with multiplicity n,p, respectively, define $G^{\lambda\mu}$ to be the $n \times p$ matrix whose k,ℓ entry is

$$-\frac{1}{1-\mu\bar{\lambda}}\left(\frac{1-|\mu|^2}{1-\mu\bar{\lambda}}\right)^k (-\bar{\mu})^{\ell-k}\binom{\ell}{k}\sum_{r=0}^{\ell-k}(-1)^r\binom{\ell-k}{r}\sum_{n=r}^{k+r}\binom{k+n}{k+r}(\delta_\lambda(\mu)\bar{\lambda})^n\ .$$

for $k \leq \ell$ and for $k \leq \ell$ it is the complex conjugate of this quantity with μ and λ permuted. Let $G^{\lambda\mu} = \{(e_k^\lambda, e_\ell^\mu)\}_{k,\ell}$ and let $\mathfrak{J}_\lambda$ be the $n \times n$ Toeplitz matrix

$$\mathcal{J}_\lambda = \begin{pmatrix} b_0(\lambda) & b_1(\lambda) & & b_{n-1}(\lambda) \\ 0 & b_0(\lambda) & b_1(\lambda) & \\ 0 & 0 & b_0(\lambda) & \ddots \\ 0 & 0 & 0 & \ddots\ b_0(\lambda) \end{pmatrix},$$

and define an $n \times n$ matrix for each pair i,j:

$$M_{ij} = G^{\lambda_i\lambda_j} - \mathcal{J}^*_{\lambda_i} G^{\lambda_i\lambda_j} \mathcal{J}_{\lambda_j} .$$

Then the matrix whose entries are M_{ij} is $\Lambda_{T_S}(\psi)$, that is,

$$\Lambda_{T_S}(\psi) = \{M_{ij}\}_{i,j} .$$

THEOREM I. There is a Σ in the $\mathcal{G}$ orbit of S satisfying $|\Sigma(e^{i\theta})| \le \mu(e^{i\theta})$ for all θ if and only if some Σ in the $\mathcal{G}$ orbit of S satisfies

$$|\Sigma(e^{i\theta})| \equiv \mu(e^{i\theta}) ,$$

and this occurs if and only if $\Lambda_{T_S}\left(\frac{S_c}{\alpha}\right)$ is a non-negative definite matrix.

If one were to (Cayley) transform this result from Δ to the R.H.P., then it would have the same form but we would use the transformed function S'_c instead of S_c, μ' instead of μ, etc. Also there is a matrix Λ'_{T_S} which arises from Λ_{T_S}. Then physically the theorem says (disregarding units and *constants of proportionality):*

There exists a matching circuit U *which delivers power to the load* S *equaling* $1-\mu'(j\omega)^2$ *or more at each frequency* ω *if and only if a matching circuit exists delivering precisely that power for all* ω *and this occurs if and only if* $\Lambda'_{T_S}\left(\frac{S'_c}{\alpha'}\right)$ *is non-negative definite.*

We emphasize that this test could be easily programmed and quickly executed for any reasonable situation with strictly passive S. In fact, if $1-\mu^2$ depends on a parameter (as it frequently does) admissible values of the parameter could be found readily by a Newton's method, since the main step is solution of a generalized $A = \lambda B$ eigenvalue problem.

MATHEMATICAL TOOLS: THE COMMUTANT LIFTING THEOREM

Let $L^2(\mathbb{C}^n)$ denote the square integrable $\mathbb{C}^n$-valued functions and let $H^2(\mathbb{C}^2)$ be the subspace consisting of functions with an analytic continuation onto Δ. If $F \in L^\infty(\mathbb{C}^n)$ let $\mathfrak{m}_F$ denote the operator $[\mathfrak{m}_F g](e^{i\theta}) \equiv F(e^{i\theta})g(e^{i\theta})$ acting on functions in $L^2(\mathbb{C}^n)$. Subspaces of the form

$$\mathcal{S}(A) \triangleq H^2 \ominus A H^2 ,$$

where $A \in H^\infty(\mathbb{C}^n)$ play a big role in operator theory. Also *compression operators* are ones of the form

$$\Gamma_A(F) \triangleq P_{H^2}\mathfrak{m}_{F*}\Big|_{\mathcal{S}(A)} ,$$

where P_A denotes the orthogonal projection of L^2 onto $\mathcal{S}(A)$ and P_{H^2} onto H^2. Consider the convex set

$$\Omega_A(F) = \{G \in H : \Gamma_A(G) = \Gamma_A(F)\} .$$

An extremely general interpolation problem first studied by D. Sarason [S] is: Find the G in Ω_F with smallest sup norm, i.e., smallest $\rho_1(G)$. Nagy and Foias subsequently answered the question in full generality [N-F] [D-M-P]. While Adamajan-Arov-Krein [A-A-K1,2,3] virtually wrote down all solutions to the problem. The Sarason-Nagy-Foias result called the commutant lifting theorem is

THEOREM. For $F \in H^\infty(\mathbb{C}^n)$ the smallest $G \in \Omega_F$ has

$$\sup_\theta \|G(e^{i\theta})\|_{\mathbb{C}^n} = \text{norm of } \Gamma_A(F) \text{ acting as an operator on } L^2(\mathbb{C}^n) .$$

Our matching result follows from this, refinements of it, and a nontrivial computational reduction.

GENERAL THEOREM

We call a function F in $H^\infty(\mathbb{C}^n)$ inner if its boundary values $F(e^{i\theta})$ are unitary a.e. Now to a given strictly passive matrix valued S we associate several functions. First denote by A_S the right spectral factor of $1 - S(e^{i\theta})^*S(e^{i\theta})$ satisfying $A_S^{-1} \in H^\infty$ and by B_S' the left spectral factor of $1 - S(e^{i\theta})S(e^{i\theta})^*$ satisfying $B_S'^{-1} \in H^\infty$. The function B_S' has a unique (up to unitary multiple) factorization $B_S' = B_S V$ with B_S, B_S^{-1} H^∞ and V an inner function (c.f., Ch. V. §4, paragraph 3 [N-F]). The function

$$\not{S}_c \triangleq - A_S S^*(1-SS^*)^{-1}B_S$$

equals $-A_S(1-S^*S)^{-1/2}S^*(1-SS^*)^{-1/2}B_S$ and so is in $L^\infty(\mathbb{C}^n)$, however, it is frequently not in $H^\infty(\mathbb{C}^n)$. By the Beurling theorem there is a unique inner function U_S such that $U_S \not{S}_c$ is in $H^\infty(\mathbb{C}^n)$, which is minimal in the usual sense that if $V \not{S}_c$ is in $H^\infty(\mathbb{C}^n)$ there is an inner W such that $V = WU_S$. Define the *complementary function* S_c of S to be $U_S \not{S}_c$.

THEOREM II. There is a Σ in the $\mathfrak{G}$ orbit of S satisfying $\|\Sigma(e^{i\theta})\| \leq \mu(e^{i\theta})$ if and only if some Σ in the orbit of S satisfies $\|\Sigma(e^{i\theta})\| \equiv \mu(e^{i\theta})$ and this occurs if and only if $\left\|\Gamma_{U_S}\left(\frac{S_c}{\alpha}\right)\right\| \leq 1$.

For $S \in H^\infty(\mathbb{C}^1)$ Theorem I and Theorem II are equivalent. The two definitions of S_c turn out to be the same and because one can use Sarason's original observation that certain explicit functions form a basis for $\mathcal{S}(U_S)$ consisting of generalized eigenvectors for $\Gamma_{U_S}\left(\frac{S_c}{\alpha}\right)$. This allows us to convert the $\|\ \| \leq 1$ condition to the very explicit matrix condition in Theorem I. Although I have not worked it through I certainly feel that a straightforward, though grubby application of Sarason's observation would make the $S \in H^\infty(\mathbb{C}^n)$ case explicit at least for small n. Here the computation of U_S can probably be bypassed, while the computation of S_c might well be the most difficult practical problem.

NON-STRICTLY PASSIVE LOADS

Now we turn from strictly passive S to ones which are merely passive. Define $\bar{\mathcal{B}}$ like $\mathcal{B}$ but only require $\|\ \| \leq 1$. If $S \in \bar{\mathcal{B}}$ but is not inner its transmission zeroes T_S do not lie strictly interior to $|z| < 1$. As mentioned the commutant lifting theorem amounts to a unified treatment of classical interpolation at points interior to the disk. In the matching problem what one must be able to do is interpolation at points of T_S. Thus for strictly passive S the commutant lifting theorem works perfectly. Unfortunately, it does not address boundary interpolation.

There are several mathematically natural interpolation problems at points on the circle. They all turn out to be trivial except for one; that is the classical Lowner interpolation problem. Ironically enough that is the one which comes up in the matching problem. Thus what is needed is a combined P-N-C-F-Lowner, etc. interpolation theory. None has been written out in detail although the principles involved are well understood and instruct one in how to compute anything he has the endurance to compute (see [D]).

With this tool *one can compute gain-bandwidth limitations as in Theorem I for any* $S \in \bar{\mathcal{B}}$ *which has at most order* 2 *transmission zeroes on* $|z| = 1$. Although we don't state this result it can be found in §6 of [H].

I have worked out the gain-bandwidth test for some cases with higher order boundary transmission zeroes. In particular, consider the R.H.P. instead of the disk for a moment. Matching problems which some engineers in the microwave amplifier industry find of interest involve loads with high order trasmission zeroes at 0 and ∞. So I worked out the gain-bandwidth n this case using the 'method of divided differences' (see [D] or [H] for a interpretation). An interested reader is welcome to write me for a copy. lso a longer discussion of this general subject appears in §6 [H].

COMPATIBLE IMPEDANCES: ORBIT STRUCTURE

The talk has not been aimed at the problem of compatible impedances but since the result for $\mathcal{B}(\mathbb{C}^1)$ is easy to state we might as well do that.

THEOREM III. If Σ, S are rational functions in $\mathcal{B}(\mathbb{C}^1)$, then Σ is in the orbit of S under $\mathcal{G}$ if and only if

$$T_S \subset T_\Sigma \qquad [\text{zeroes of } \Sigma_c] \cap T_S = \phi$$

$$\Lambda_{T_S}\left(\frac{S_c}{\Sigma_c}\right) \geq 0 .$$

HISTORY

The first work on matching and gain bandwidth limitations was done by Bode in 1947 on a very simple load circuit. Fano in the early 1950's did a more general study. The prevailing theoretical method was introduced by Youla in 1967. It deals entirely with one ports and in the literature the one port matching problem is frequently referred to as a solved problem. The recipe that is given, however, involves guessing primarily as to gain-bandwidth limitations. The procedure recommended for divining gain-bandwidth limitations uses 'Fano's-integral constraints'. It is always presented via a series of (simple) examples rather than explicitly and I would guess it becomes very awkward when order $T_S > 4$. The most concrete method for determining the limitations is in Problem 2 [Y-S] or Theorem 4.2 [C]. It holds for S with order one transmission zeroes and $\mu = 1$. We note here that the constructions in [H] and the mathematical work it relies on might present useful alternatives to the existing recipe. For n-ports theory is very incomplete with the most advanced work being that of R. Wohlers [W].

Textbook accounts of broadband matching theory are [C], [K-R], Ch. 3, [W]. Also in this came conderence proceedings there are four papers resulting from a special session on broadband matching.

ACKNOWLEDGEMENT

Thanks are certainly due to Neola Crimmins for her careful typing of the hastily prepared and almost illegible manuscript.

REFERENCES

[C] Wai-Kai Chen, *Theory and Design of Broadband Matching Networks,* Pergamon Press, New York, 1976.

[D] W. F. Donoghue, *Monotone Matrix Functions and Analytic Continuation,* New York, Springer, 1974.

[H] J. W. Helton, *Orbit Structure of the Mobius Transformation Semigroup Acting on H^∞ (Broadband Matching),* preprint available.

[K-R] E. Kuh and R. A. Roher, *Theory of Linear Active Networks,* San Francisco, Holden-Day, 1967.

[W] R. Wohlers, *Lumped and Distributed Passive Networks,* Academic Press, New York, 1969.

[Y-S] D. C. Youla and M. Saito, *Interpolation with Positive Real Functions,* J. Franklin Institute, vol. 284, no. 2, 1967, pp. 77-108.

VULNERABILITY OF DYNAMIC SYSTEMS[*]

D. D. Šiljak
School of Engineering
University of Santa Clara
Santa Clara, California

ABSTRACT

Directed graphs are associated with dynamic systems in order to determine in any given system if each state can be reached by at least one input (input reachability), or can each state reach at least one output (output reachability). Then, the structural perturbations of a dynamic system are identified as lines or points removals from the corresponding digraph, and a system is considered vulnerable at those lines or points of the digraph whose removal destroys its input or output reachability. A suitable framework is formulated for resolving the problems of reachability and vulnerability which applies to both linear and nonlinear systems alike.

I. INTRODUCTION

Numerous natural, social, and technological systems which represent a dynamic interaction among a number of elements or subsystems, quite commonly either by design or fault, do not stay "in one piece" during operation. They are often subject to *structural perturbations* whereby groups of subsystems are disconnected from and again connected to each other in an unpredictable way. The effects of such perturbations on stability of various models in ecology, economics, and engineering were explored in a number of studies [1] where it was shown that a class of stable competitive systems is invulnerable to perturbations in the interconnection structure.

A general framework for studying the pure structural properties of dynamic systems was proposed in reference [2], where the notions of *input* and *output reachability* were introduced in terms of digraphs. A system is input-reachable if each state can be reached by at least one input, and it is output reachable if each state can reach at least one output. Once a digraph is associated with a dynamic system and input (output) reachability is defined in digraph theoretic terms, we formulate the vulnerability problem as follows: Does a removal of a line or a point from the digraph destroy input (output) reachability? In proving an answer to this question, it will be shown that input (output) decentralized systems are a structural counterpart of stable competitive systems in that they are invulnerable under structural perturbations.

The plan of the paper is the following:

In the next section, we fix the relationship between dynamic systems, directed graphs (digraphs) and interconnection matrices. Inputs, states, and outputs of a system are represented by input, state, and output points of the corresponding digraph. The lines of the digraph and the entries of the related interconnection matrix are determined in the usual way by occurrence of inputs, states, and outputs in the system equations.

In Section 3, we introduce the notion of input and output truncated digraphs, and use the reachability and antecedent sets to define *input* and *output reachability* of dynamic systems. Then, using the path matrix we develope an algebraic criterion for input and output reachability which can be applied to dynamic systems using effective algorithms already available in digraph studies of computer systems.

---

[*] The research presented herein was supported by NASA, Ames Research Center, Moffett Field, California, under the Grant NGR 05-017-010.

Section 4 is devoted to *partitions* and *condensations* of digraphs as related to dynamic systems, which provide an appropriate setting for analyzing the important reachability properties of large-scale systems composed of interconnected subsystems. In this context, we will formulate a *pure canonical structure* of nonlinear systems with respect to input and output reachability.

In Section 5 we formulate the notion of *structural perturbations* of composite dynamic systems as line removals and point removals from the corresponding condensations. Then, we explore the conditions under which input (output) reachable dynamic systems are invulnerable to the structural perturbations.

Finally, in Section 6, we make use of partitions and condensations of digraphs to define input and output *decentralized systems*. A somewhat trivial but nevertheless important result is that the decentralized systems remain input (output) reachable despite structural perturbations whereby subsystems are disconnected and again connected in various ways during the functioning of the overall system.

II. BASIC STRUCTURE

Let us consider a system $\mathcal{S}$ which is described by the equations

$$\dot{x} = f(t, x, u)$$
$$y = g(t, x) \ , \tag{1}$$

where $x(t) \in R^n$ is the state, $u(t) \in R^m$ is the input, and $y(t) \in R^\ell$ is the output of $\mathcal{S}$. The functions $f\colon R \times R^n \times R^m \to R^n$ and $g\colon R \times R^n \to R^\ell$ are sufficiently smooth so that $\mathcal{S}$ represents a *dynamic system* [3].

With the dynamic system $\mathcal{S}$ we associate a *directed graph (digraph)* [4] as the ordered pair $\mathcal{D} = (V, R)$ where $V = U \cup X \cup Y$ and $U = \{u_1, u_2, \ldots, u_m\}$, $X = \{x_1, x_2, \ldots, x_n\}$, $Y = \{y_1, y_2, \ldots, y_\ell\}$ are nonempty sets of input, state and output points, respectively. R is a relation in V, that is, R is a set of ordered pairs which are the lines (u_j, x_i), (x_j, x_i), or (x_j, y_i) joining the points of $\mathcal{D}$. We make an important assumption about $\mathcal{S}$ by requiring that $\mathcal{D}$ does not contain lines of the type (u_j, u_i), (x_j, u_i), (y_j, x_i), (y_j, u_i), and (y_j, y_i). This requirement may seem to be overly restrictive, but in fact it is not since it reflects the structure of what we ordinarily consider as a dynamic system $\mathcal{S}$ described by equation (1). We merely assume that there are no lines joining the input points, no lines from the input points to the output points, etc.

A convenient way to represent a digraph $\mathcal{D}$ associated with $\mathcal{S}$, is to use interconnection matrices. We propose that the $p \times p$ interconnection matrix $M = (m_{ij})$ which we define as a composite matrix

$$M = \left[\begin{array}{c|c|c} E & L & 0 \\ \hline 0 & 0 & 0 \\ \hline F & 0 & 0 \end{array}\right] , \tag{2}$$

such that $i, j = 1, 2, \ldots, p$ and $p = n + m + \ell$, be used to describe the *basic structure of* $\mathcal{S}$. In (2), the $n \times n$ state connection matrix $E = (e_{ij})$ is defined as a binary matrix with elements e_{ij} specified by

$$e_{ij} = \begin{cases} 1 \ , & (x_j, x_i) \in R \\ 0 \ , & (x_j, x_i) \notin R \ , \end{cases} \tag{3}$$

where $i, j = 1, 2, \ldots, n$. That is, $e_{ij} = 1$ if x_j "occurs" in $f_i(t, x, u)$, and $e_{ij} = 0$ if x_j "does not occur" in $f_i(t, x, u)$. Similarly, we define the $n \times m$ input connection matrix $L = (\ell_{ij})$ as

$$\ell_{ij} = \begin{cases} 1\,, & (u_j, x_i) \in R \\ 0\,, & (u_j, x_i) \notin R\,, \end{cases} \tag{4}$$

where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. In other words, $\ell_{ij} = 1$ if u_j occurs in $f_i(t, x, u)$, and $\ell_{ij} = 0$ if u_j does not occur in $f_i(t, x, u)$. Finally, the $\ell \times n$ output connection matrix $F = (f_{ij})$ is defined by

$$f_{ij} = \begin{cases} 1\,, & (x_j, y_i) \in R \\ 0\,, & (x_j, y_i) \notin R\,, \end{cases} \tag{5}$$

where $i = 1, 2, \ldots, \ell$ and $j = 1, 2, \ldots, n$. Again, $f_{ij} = 1$ if x_j occurs in $g_i(t, x)$, and $f_{ij} = 0$ if x_j does not occur in $g_i(t, x)$.

III. INPUT AND OUTPUT REACHABILITY

If each state x_i can be reached by at least one input u_j either directly or indirectly through other states of the system S, then we say that the system S is *input reachable*. When all states x_j reach at least one output y_i either directly or via other states of S, then each x_j is represented by the output of S. In this case we say that the system S is *output reachable*. If each input u_j reaches at least one output y_i and each y_i is reached by at least one u_j, we say that S is *input-output reachable*.

To express precisely the reachability properties of S in graph theoretic terms we recall [4] the notions of the *reachable* and *antecedent sets* of $\mathcal{D}$, and state the following [2]:

Definition 1. A system S with a digraph $\mathcal{D} = (U \cup X \cup Y, R)$ is input reachable if and only if X is the reachable set of U.

The "directional dual" of Definition 1 is:

Definition 2. A system S with a digraph $\mathcal{D} = (U \cup X \cup Y, R)$ is output reachable if and only if X is the antecedent set of Y.

By imitating Definitions 1 and 2, but otherwise using different sets, we can formulate the following:

Definition 3. A system S with a digraph $\mathcal{D} = (U \cup X \cup Y, R)$ is input-output reachable if and only if Y is both the reachable set and the antecedent set of U.

From the above definitions, we can conclude that if a system S is both input and output reachable it is also input-output reachable, but the converse is not true in general.

Another characterization of input and output reachability is possible by using the notion of *point basis* [4] and the definitions of *input* and *output truncations* introduced in [2]:

Definition 4. The input truncation $\mathcal{D}_u = (U \cup X, R_u)$ of a digraph $\mathcal{D} = (U \cup X \cup Y, R)$ is the maximal subgraph of $\mathcal{D}$ not containing points of Y.

In other words, $\mathcal{D}_u$ has as its set of points $U \cup Y$, and has as its set of lines R_u all lines of $\mathcal{D}$ except those lines incident with points of Y.

Similarly, we have:

Definition 5. The output truncation $\mathcal{D}_y = (X \cup Y, R_y)$ of a digraph $\mathcal{D} = (X \cup Y, R)$ is the maximal subgraph of $\mathcal{D}$ not containing points of U.

By combining Definitions 4 and 5, we obtain the definition of the *input-output truncation* $\mathcal{D}_x = (X, R_x)$ as a digraph which has as its set of points

X , and has as its set of lines R_X all lines of $\mathcal{D}$ except those lines incident with the points of U and Y .

We prove the following:

Theorem 1. A system S with a digraph $\mathcal{D}$ is input reachable if and only if U is a point basis of the input truncation $\mathcal{D}_u$.

Proof. To see the "if" part we recall that if U is the point basis of $\mathcal{D}$, then U is a (minimal) collection of points from which all of its points, and thus points of X , are reachable. Conversely, if U is not the point basis of $\mathcal{D}_u$ then there would be points of X which cannot be reached from points of U and S is not input reachable. This proves Theorem 1.

Similarly, we have:

Theorem 2. A system S with a digraph $\mathcal{D}$ is output reachable if and only if Y is the point contrabasis of the output truncation $\mathcal{D}_y$.

Proof. Theorem 2 is a dual of Theorem 1.

Theorems 1 and 2 can be combined to provide a characterization of input-output reachable dynamic systems in terms of point basis.

For algorithmic purposes, it is convenient to consider reachability of S in terms of $p \times p$ *path matrix* $P = (p_{ij})$ which corresponds to the digraph $\mathcal{D} = (V, R)$ and is defined as

$$p_{ij} = \begin{cases} 1 , \text{ there is a path from } v_j \text{ to } v_i \\ 0 , \text{ there is no path from } v_j \text{ to } v_i . \end{cases} \tag{6}$$

In definition (6) of P , the trivial paths of zero length are excluded; the length of a path being the number of lines in the path.

To determine the path matrix for a given digraph, we need the following result obtained by Festinger et al. [5]:

Theorem 3. Let $M = (m_{ij})$ be the $p \times p$ interconnection matrix corresponding to a digraph $\mathcal{D} = (V, R)$, and let $N = (n_{ij})$ be the $p \times p$ matrix such that $N = M^d$ where $d \in \{1, 2, \ldots , p\}$. Then, n_{ij} is the total number of distinct sequences $(v_j, \ldots), \ldots , (\ldots , v_i)$ of length d in $\mathcal{D}$.

Proof. We can prove this theorem by induction following Berztiss [6]. We show that the theorem is true for $d = 1$, and then show that it is true for $d + 1$ whenever it is true for d . Then, the theorem is true for any d . For $d = 1$, the theorem is actually the definition of the matrix M . For $d + 1$, $m_{ik}n_{kj} = n_{kj}$ if (v_k, v_i) is a line, and $m_{ik}n_{kj} = 0$ if (v_k, v_i) is not a line. The total number of sequences of length $d + 1$ having the form $(v_j, \ldots), \ldots , (\ldots , v_k), (v_k, v_j)$ is equal to $\sum_{k=1}^{p} m_{ik}n_{kj}$, which is the ij-th element of M^{d+1} . This proves Theorem 3.

The path matrix is calculated using the following:

Corollary 1. Let $P = (p_{ij})$ be a path matrix of $\mathcal{D} = (V, R)$, and $Q = (q_{ij})$ be a matrix defined as

$$Q = M + M^2 + \ldots + M^p . \tag{7}$$

Then, $p_{ij} = 1$ if and only if $q_{ij} \neq 0$.

Once reachability is formulated in terms of the path matrix P , we can calculate P from a given M to determine the input and output reachability of S .

Let us note that M^d can be written as

$$M^d = \left[\begin{array}{c:c:c} E^d & E^{d-1}L & 0 \\ \hdashline 0 & 0 & 0 \\ \hdashline F\,E^{d-1} & F\,E^{d-2}L & 0 \end{array}\right] , \tag{8}$$

where again $d \in \{1, 2, \dots, p\}$. The matrix Q of (7) has the form

$$Q = \left[\begin{array}{c|c|c} A_0 & B_0 & 0 \\ \hline 0 & 0 & 0 \\ \hline C_0 & D_0 & 0 \end{array}\right] \tag{9}$$

where

$$\begin{aligned} A_0 &= E + E^2 + \dots + E^p \\ B_0 &= (I + E + \dots + E^{p-1})L \\ C_0 &= F(I + E + \dots + E^{p-1}) \\ D_0 &= F(I + E + \dots + E^{p-2})L \ . \end{aligned} \tag{10}$$

We arrive immediately at the following [2]:

Theorem 4. A system S *with an input-output connection matrix* M *defined in* (2) *is input reachable if and only if the matrix* B_0 *of* (10) *has no zero rows, it is output reachable if and only if the matrix* C_0 *of* (10) *has no zero columns, and it is input-output reachable if and only if* D_0 *of* (10) *has neither zero rows nor zero columns.*

Proof. By constructing the path matrix P using Q of (9) and Corollary 1, the proof of Theorem 4 is automatic.

The computation of the path matrix by generating powers of the interconnection matrix as performed above, is not a numerically attractive procedure for verifying reachability properties of the corresponding dynamic system. There are numerous algorithms developed to avoid various numerical difficulties in identifying the paths of a digraph, which started by the well-known Boolean representation algorithm of Warshall [7] and culminated by the depth-first search method of Tarjan [8]. A survey of these algorithms is given by Bowie [9].

IV. PARTITIONS AND CONDENSATIONS

For either conceptual or numerical reasons, large-scale dynamic models in ecology [10], economics [11], and engineering [12], are considered as dynamic systems partitioned into interconnected subsystems. In order to gain insight into the structure of such models, we propose to investigate their *condensation digraphs* [4]. These digraphs are obtained from the original ones by replacing the subgraphs corresponding to the subsystems by points, and joining the new points by lines which represent the interconnections among the subsystems.

Let us consider again a system S described by the equations

$$\begin{aligned} \dot{x} &= f(t, x, u) \\ y &= g(t, x) \ , \end{aligned} \tag{1}$$

where $x(t) \in R^n$ is the state, $u(t) \in R^m$ is the input, and $y(t) \in R^\ell$ is the output of S. We assume that S is decomposed into s interconnected subsystems S_i described by the equations

$$\begin{aligned} \dot{x}_i &= f_i(t, x, u) \\ y_i &= g_i(t, x), \ i = 1, 2, \dots, s \end{aligned} \tag{11}$$

where $x_i(t) \in R^{n_i}$ is the state, $u_i(t) \in R^{m_i}$ is the input, and $y_i \in R^{\ell_i}$ is the output of S_i. We have

$$\begin{aligned} R^n &= R^{n_1} \times R^{n_2} \times \dots \times R^{n_s} \\ R^m &= R^{m_1} \times R^{m_2} \times \dots \times R^{m_r} \\ R^\ell &= R^{\ell_1} \times R^{\ell_2} \times \dots \times R^{\ell_q} \end{aligned} \tag{12}$$

so that

$$x = (x_1^T, x_2^T, \dots, x_s^T)^T$$
$$u = (u_1^T, u_2^T, \dots, u_r^T)^T$$
$$y = (y_1^T, y_2^T, \dots, y_q^T)^T \tag{13}$$

are the state, input and output vectors of S, respectively.

We associate again a digraph $\mathcal{D} = (V, R)$ with the dynamic system S described by (1). Then we partition each set U, X, Y of V into disjoint subsets $u_1, u_2, \dots, u_r$; $x_1, x_2, \dots, x_s$; $y_1, y_2, \dots, y_q$ whose union is all of V. Thus, each element of V is in exactly one subset. The condensation $\mathcal{D}^* = (V^*, R^*)$ of $\mathcal{D} = (V, R)$ with respect to this partition is the digraph whose points V^* are those subsets of V^*, that is, $V = U^* \cup X^* \cup Y^*$ and $U^* = \{u_1, u_2, \dots, u_r\}$, $X^* = \{x_1, x_2, \dots, x_s\}$, $Y^* = \{y_1, y_2, \dots, y_q\}$. The lines of the condensation $\mathcal{D}^*$ are determined by the following rule: There is a line (u_j, x_i) in R^* if and only if there is at least one line in R from a point of the subset u_j of U to a point of the subset x_i of X. Similarly, the rule holds for the lines (x_j, x_i) and (x_j, y_i) of $\mathcal{D}^*$. In this way, the condensation $\mathcal{D}^*$ represents (uniquely) the structure of the composite system S described by (11) with points of $\mathcal{D}^*$ standing for the subsystems and the lines of $\mathcal{D}^*$ standing for the interconnections among them.

Another way to represent a partition of S is to use interconnection matrices in the same way as they were used to describe the original system S. Rewriting the matrix M of (2), but otherwise using different submatrices, we can define the $p \times p$ matrix

$$M^* = \left[\begin{array}{c|c|c} E^* & L^* & 0 \\ \hline 0 & 0 & 0 \\ \hline F^* & 0 & 0 \end{array}\right] \tag{14}$$

which we associate with the condensation $\mathcal{D}^*$ in an obvious way. In (14), E^*, L^*, F^* are $s \times s$, $s \times r$, $q \times s$ matrices, respectively, and $p = s + r + q$. Now, input and output reachability of the condensation $\mathcal{D}^*$ can be determined by applying Theorem 2 to the matrix M^* of (14).

There are many different ways in which a dynamic system S and the corresponding digraph $\mathcal{D}$ may be partitioned. In the pure theory of structures [4, 13-15] it is common to partition a digraph into its *strong components* so that each point of the condensation $\mathcal{D}^*$ corresponds to one and only one strong component of $\mathcal{D}$; a strong component of a digraph $\mathcal{D}$ being a subgraph of $\mathcal{D}$ in which every two points are mutually reachable. In order to conclude input and output reachability of such special condensation, we denote by $\mathcal{D}_x^*$ the truncated condensation of $\mathcal{D}^*$, which corresponds to the truncation $\mathcal{D}_x$, and prove the following:

Theorem 5. Let the condensation $\mathcal{D}_x^*$ be constructed with respect to the strong components of $\mathcal{D}_x$. Then, the digraph $\mathcal{D}$ is input (output) reachable if and only if the condensation $\mathcal{D}^*$ is input (output) reachable.

Proof. We prove only the input reachability part of the theorem since the output reachability part is its directional dual. If $\mathcal{D}^*$ is input reachable than X^* is the reachable set of U^* and there is a path to each point of X^* from a point of U^*. Since $\mathcal{D}_x^*$ is a condensation with respect to strong components of $\mathcal{D}_x$, it follows from reachability of the components that there is a path to each point of X from a point of U. That is, X is the reachable set of U and $\mathcal{D}$ is input reachable. Conversely, if $\mathcal{D}^*$ is not input reachable, then X^* is not the reachable set of U^*, and there are points of X^* that cannot be reached by any point of U^*. Obviously, by the definition of condensation, those points of X that correspond to the unreachable points of X^* cannot be reached by any point of U, and $\mathcal{D}$ is not input reachable. This proves Theorem 5.

With an abuse of Definitions 1 and 2, in the above proof we refer to input (output) reachability of the digraphs $\mathcal{D}$ and $\mathcal{D}^*$. This was done to avoid ambiguity arising from the fact that both the digraph $\mathcal{D}$ and its condensation $\mathcal{D}^*$ are related to the same dynamic system $\mathcal{S}$. The change in terminology should create no confusion since input (output) reachability is defined unambiguously in terms of digraphs. Still another abuse of notation is committed in (13), where it would be appropriate to use starred notation on all components of the vectors x, u, and y since they represent a partition and condensation rather than the individual states, inputs, and outputs of the system $\mathcal{S}$. This again should not pose any difficulty since in the rest of the paper, if not indicated to the contrary, all components of state, input, and output vectors correspond to partitions and condensations.

Algorithms of Purdom [13], Munro [14], and Kevkorian [15] can be used to compute the strongly connected components of the digraph $\mathcal{D}$. Then, input and output reachability can be determined by forming the condensation $\mathcal{D}_x^*$ and applying Theorem 5.

Partitions of mathematical models of dynamic processes in ecology [10], economics [11], and engineering [12], are most often guided by the special structural properties of the models. Therefore, in general, partitions of the digraphs are not performed with respect to their strong components, and there is no reason why the corresponding condensations have to be constituted that way. For example, for reasons of control and estimation [16-18], it is desirable to partition a dynamic system $\mathcal{S}$ into input and output reachable subsystems $\mathcal{S}_i$. That is, it is advantageous to partition the corresponding digraph $\mathcal{D}$ into input and output reachable components which are not necessarily strong. This is considered next.

In his controllability and observability considerations, Kalman [19] formulated a canonical structure of linear dynamic systems. By choosing appropriately the coordinate system, he showed that each linear system can be partitioned into four parts depending on whether controllability and observability hold or not. By imitating such development, but otherwise using input and output reachability, we can formulate the *pure canonical structure* of dynamic systems.

Let us consider again the dynamic system $\mathcal{S}$ described by (1) together with its interconnection matrix M of (2). By permutation of rows and columns of M, it can be transformed into a matrix $\tilde{M}$ which has the following form:

$$\tilde{M} = \begin{array}{c} \left[\begin{array}{cccc|c|c} E_{11} & E_{12} & 0 & 0 & L_1 & 0 \\ 0 & E_{22} & 0 & 0 & 0 & 0 \\ E_{31} & E_{32} & E_{33} & E_{34} & L_3 & 0 \\ 0 & E_{42} & 0 & E_{44} & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ \hline F_1 & F_2 & 0 & 0 & 0 & 0 \end{array}\right] \begin{array}{l} \tilde{x}_{io} \\ \tilde{x}_{\not{i}o} \\ \tilde{x}_{i\not{o}} \\ \tilde{x}_{\not{i}\not{o}} \\ u \\ y \end{array} \\ \begin{array}{cccccc} \tilde{x}_{io} & \tilde{x}_{\not{i}o} & \tilde{x}_{i\not{o}} & \tilde{x}_{\not{i}\not{o}} & u & y \end{array} \end{array} \tag{15}$$

The partition of the state vector of the transformed system $\tilde{\mathcal{S}}$,

$$\tilde{x} = (\tilde{x}_{io}^T, \tilde{x}_{\not{i}o}^T, \tilde{x}_{i\not{o}}^T, \tilde{x}_{\not{i}\not{o}}^T)^T, \tag{16}$$

into four components, which can be identified from $\tilde{M}$ in (15), represent the four subsystems with the following properties:

$(\tilde{S}_{io})$ input reachable and output reachable
$(\tilde{S}_{\not{i}o})$ input unreachable and output reachable
$(\tilde{S}_{i\not{o}})$ input reachable and output unreachable
$(\tilde{S}_{\not{i}\not{o}})$ input unreachable and output unreachable.

By using condensation, we can represent each component of x in (16) as a point of the condensation $\tilde{\mathcal{D}}^*$ as shown in Figure 1. This digraph represents the pure canonical structure of the dynamic system S.

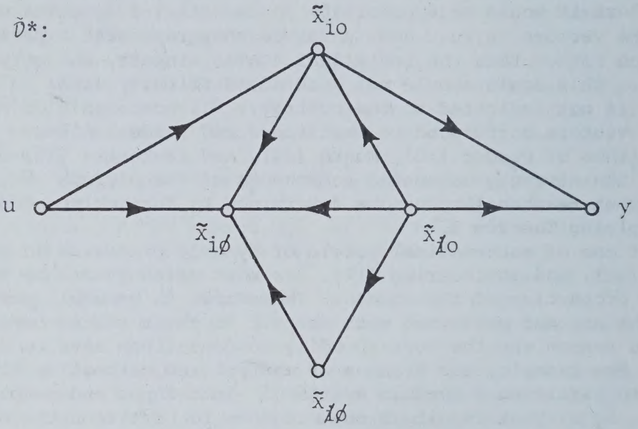


Fig. 1. Canonical Structure

There are as least three good reasons why input-output reachability is more appealing than its controllability-observability counterpart. First, the input-output reachability concept applies to linear and nonlinear systems alike. Second, the difference between the rank condition for controllability and observability [19] and the connectedness test for input and output reachability can be dismissed on the physical grounds as unrealistic since a slight perturbations of parameters in input and output reachable systems can restore the controllability and observability properties [20]. Third, the rank conditions are ill-posed numerical problems as compared to structural manipulations offered by connectedness and reachability considerations [21].

Besides the canonical structure, there are other types of structures that are of interest in control systems. For estimation and control of large-scale systems [16-18], input and output decentralized models have important applications. They are considered in Section 6.

Finally, we should also mention a possibility to alter our notion of subsystems and allow "overlapping" of subspaces R^{n_i} in (12). For example, the state vector $\tilde{x}_k$ of the subsystem $\tilde{S}_k$ can be formed pair-wise as $\tilde{x}_k = (x_i^T, x_j^T)^T$ where x_i and x_j are defined in (13). This kind of partition was used in the stability analysis of interconnected power systems presented in reference [22].

V. VULNERABILITY

Suppose that a system S is input (output) reachable. And suppose furthermore, that we disconnect a number of subsystems from each other. Is input (output) reachability still present or is it destroyed by the disconnection? This is the problem of vulnerability of dynamic systems which we propose to consider in the framework of digraph theory.

In terms of digraphs [4], disconnecting subsystems from each other is equivalent to "line removals" and disconnecting subsystems from the overall system is equivalent to "point removals". These perturbations can be conveniently described by interconnection matrices as shown in [1]. For this purpose, the $s \times s$ fundamental interconnection matrix $\bar{E}^* = (\bar{e}_{ij})$ is associated with a condensation $\bar{\mathcal{D}}^* = (U^* \cup X^* \cup Y^*, \bar{R}^*)$ of $\mathcal{S}$ in (14) as follows:

$$\bar{e}^*_{ij} = \begin{cases} 1, & (x_j, x_i) \in \bar{R}^* \\ 0, & (x_j, x_i) \notin \bar{R}^* \end{cases} . \tag{17}$$

That is, $\bar{e}^*_{ij} = 1$ if x_j occurs in $f_i(t, x, u)$, and $\bar{e}^*_{ij} = 0$ if x_j does not occur in $f_i(t, x, u)$. Now, a structural perturbation is represented by a line removal (or number of lines) of the condensation $\bar{\mathcal{D}}^* = (U^* \cup X^* \cup Y^*, \bar{R}^*)$ between points of X^*. That results in a spanning subgraph $\mathcal{D}^*$ of $\bar{\mathcal{D}}^*$, that is, a subgraph with the same set of points as $\bar{\mathcal{D}}^*$. All spanning subgraphs of $\bar{\mathcal{D}}^*$ obtained this way, can be represented uniquely by an interconnection matrix E^* which is obtained from $\bar{E}^*$ as follows: $\bar{e}_{ij} = 0$ in $\bar{E}^*$ implies $e^*_{ij} = 0$ in E^* for all $i, j = 1, 2, \ldots, s$; and removal of a line (x_j, x_i) of $\mathcal{D}^*$ implies that $\bar{e}^*_{ij} = 1$ in $\bar{E}^*$ is replaced by $e^*_{ij} = 0$ in E^*. The fact that an interconnection matrix E^* is generated in this way by the fundamental interconnection matrix $\bar{E}^*$ is denoted by $E^* \in \bar{E}^*$. Finally, without loss of generality, a point removal can be treated in the same way as a special case of line removals. If a k-th point of $\bar{\mathcal{D}}^*$ is removed, then $e^*_{ik} = e^*_{kj} = 0$ for all $i, j = 1, 2, \ldots, s$.

Structural perturbations are illustrated in Figure 2. The top digraph $\bar{\mathcal{D}}^*_y$ is the output truncation which represents the basic structure of the composite system $\mathcal{S}$ when all the inputs are removed, and which corresponds to the fundamental interconnection matrix $\bar{E}^*$. The structural perturbations formed by line removals are represented by the digraphs below the digraph $\bar{\mathcal{D}}^*_y$. The digraph $\mathcal{D}^*_y$ at the bottom of Figure 2 corresponds to $E^* = 0$ and, thus, the total disconnection of lines connecting the state points x_1, x_2, x_3 of the set X^*. In an obvious way, structural perturbations of systems with inputs can be visualized using the same Figure 2.

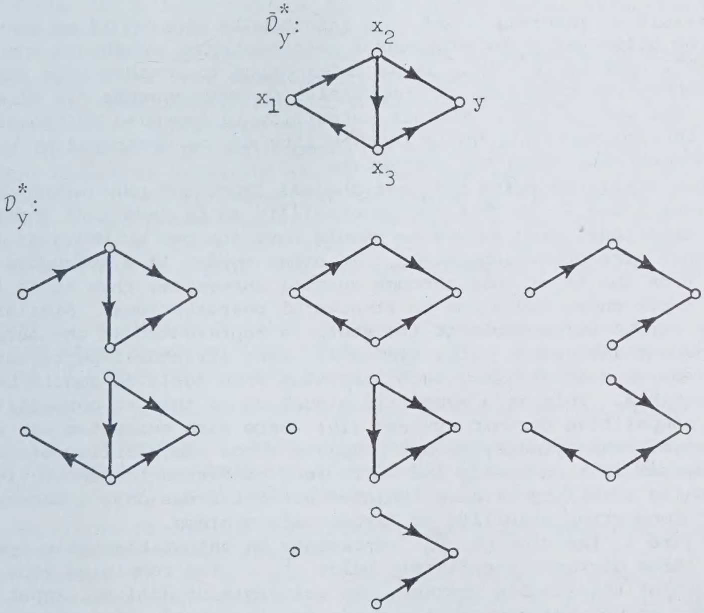


Figure 2. Structural Perturbations

On the basis of the above considerations, we introduce the following:

Definition 6. *A system* S *with a condensation* $\bar{\mathcal{D}}^* = (U^* \cup X^* \cup Y^*, \bar{R}^*)$ *is connectively input (output) reachable if and only if it is input (output) reachable for all interconnection matrices* $E^* \in \bar{E}^*$.

Definition 6 parallels those of connective stability [1], and connective suboptimality [17]. In the case of control systems, the three definitions complement each other in an important way discussed in the latter part of this section.

Now we prove the following:

Theorem 6. *A system* S *with an input truncated condensation* $\bar{\mathcal{D}}_u^* = (U^* \cup X^*, \bar{R}_u^*)$ *is connectively input reachable if and only if the set* U^* *is a 1-basis of* $\bar{\mathcal{D}}_u^*$ *for* $E^* = 0$.

Proof. The "if" part for $E^* = 0$ follows directly from the definition of the 1-basis. For $E^* \neq 0$, we get the corresponding digraph by adding lines to the one that corresponds to $E^* = 0$. But, it is obvious that there are no lines in any digraph whose addition can destroy its reachability property. For if (x_j, x_i) is any line of $\bar{\mathcal{D}}_u^*$, any path in $\bar{\mathcal{D}}_u^* - (x_j, x_i)$ is also in $\bar{\mathcal{D}}_u^*$.

On the other hand, if U^* is not a 1-basis for $E^* = 0$, then there would be points of X^* that are not reachable from U^* for all $E^* \in \bar{E}^*$, and S is not connectively input reachable. This proves the "only if" part and, thus, Theorem 6.

By principle of duality of digraphs [4], it is possible to obtain an analog to Theorem 6 for connective output reachability of S :

Theorem 7. *A system* S *with an output truncated condensation* $\bar{\mathcal{D}}_y^* = (X^* \cup Y^*, \bar{R}_y^*)$ *is connectively output reachable if and only if the set* Y^* *is a 1-contrabasis of* $\bar{\mathcal{D}}_y^*$ *for* $E^* = 0$.

Proof. The theorem is an obvious dual of Theorem 6.

The result of Theorems 6 and 7 is intuitively clear: If we want reachability to be preserved under structural perturbations, we should check the "worst case", that of $E^* = 0$. It is interesting to compare this statement with the connective stability of competitive dynamic systems [1, 6] where the worst case was $E^* = \bar{E}^*$. That is, stability is improved by reducing the number of interconnections, while reachability may be destroyed by removing the interconnections among the subsystems.

Another straightforward but nevertheless important conclusion coming from Theorems 6 and 7, is that for reachability to be invariant under structural perturbations, each subsystem should have its own or share an input and an output with other subsystems. In other words, if a subsystem is reachable from the input only through another subsystem, then it is liable to become input unreachable due to structural perturbations. Similarly, it can become output unreachable if its state is represented at the output only by its adjacent subsystem. If a system is under structural perturbations, then to preserve reachability, each subsystem when isolated should be input-output reachable. This is a symmetric situation to that of connective stability of competitive dynamic systems [16] where each subsystem was stable when isolated. The symmetry is not complete since reachability of each isolated subsystem is a necessary and sufficient condition for connective reachability, while stability of each isolated subsystem was only a necessary condition for connective stability of large-scale systems.

In Figure 2, the digraph $\bar{\mathcal{D}}_y^*$ represents an output reachable system, and so do the three digraphs immediately below $\bar{\mathcal{D}}_y^*$. The remaining four digraphs describe output unreachable systems. By principle of duality, input reachability under structural perturbations can be visualized using the same Figure 2.

Finally, one can generalize the above connective reachability concept by defining partial connective reachability with respect to a pair of interconnection matrices $(\bar{E}^*, \hat{E}^*)$, where a fixed matrix $\hat{E}^* \in \bar{E}^*$ takes the role of $E^* = 0$. Again, it is intuitively clear that if U^* is a 1-basis for $\hat{E}^*$, this being the worst case, then S is partially connectively input-state reachable for all $E^* \in \bar{E}^* - \hat{E}^*$ where by the difference $\bar{E}^* - \hat{E}^*$ we mean all interconnection matrices generated by $\bar{E}^*$ which have the unit elements corresponding to $\hat{E}^*$.

Carrying this extension a step further and concentrating on specific set of lines we can determine the following useful structural characteristic of a dynamic system S : (i) the lines (or the maximum number of lines) of $\mathcal{D}$ whose removal does not affect input (output) reachability of S ; (ii) the lines (or the minimum number of lines) of $\mathcal{D}$ which are required to preserve input (output) reachability of S ; and (iii) the lines (or the minimum number of lines) of $\mathcal{D}$ whose removal destroys input (output) reachability of S.

Furthermore, it should be noted that there is no reason why the notion of structural perturbations should not be enlarged to include the lines which connect the inputs and outputs with the states in the corresponding digraph. Of course, the above listed problems may be harder to solve when this enlarged notion is involved.

It would be of interest to select classes of dynamic systems with canonical structures with respect to the above list of characteristics. One such structure is offered by input (output) decentralized systems which are invulnerable to arbitrary structural perturbations. They are considered next.

VI. DECENTRALIZED SYSTEMS

In large dynamic systems, decentralization is an effective way to cope with complexity and gross changes in the interconnection structure and characteristics which cannot be either predicted or eliminated during operation [16-18]. Despite a fast increase of interest in decentralized systems [23], it is only recently that attempts are made to formulate an appropriate definition of what is a decentralized system [2] and devise effective decentralization schemes [18] for dynamic systems.

Whether a system is decentralized or not is essentially a matter of its structure, and therefore, decentrality of dynamic systems can be appropriately defined in terms of digraphs. Intuitively, a system S is an *input (output) decentralized system* if each subsystem S_i has its own input u_i (output y_i). In order to put this intuitive notion of input decentralization into digraph theoretic terms and at the same time avoid trivial cases which are of no interest to us, we assume that no points in the subsets U and Y of $V = U \cup X \cup Y$ are disconnected, i.e., $od(u_k) \neq 0$ for all $k = 1, 2, \dots,$ m, and $id(y_k) \neq 0$ for all $k = 1, 2, \dots,$, where *od* and *id* stand for *out degree* and *in degree*, and use the notion of *1-basis* to state the following:

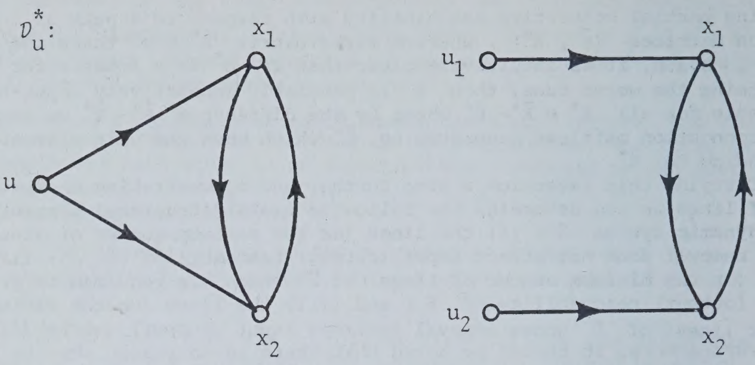
Definition 7. *A system* S *with an input truncated condensation* $\mathcal{D}_u^* = (U^* \cup X^*, R_u^*)$ *is input decentralized if and only if* $r = s$ *and the set* U^* *is a 1-basis of* $\mathcal{D}_u^*$ *such that* $od(u_i) = 1$ *for all* $i = 1, 2, \dots, r$.

By using the notion of *1-contrabasis* we state the following:

Definition 8. *A system* S *with an output truncated condensation* $\mathcal{D}_y^* = (X^* \cup Y^*, R_y^*))$ *is output decentralized if and only if* $q = s$ *and the set* Y *is a 1-contrabasis of* $\mathcal{D}_y^*$ *such that* $id(y_i) = 1$ *for all* $i = 1, 2, \dots, q$.

It is simple to combine Definitions 7 and 8 and formulate a definition of *input-output decentralized systems*.

The digraph $\mathcal{D}_u^*$ of an input centralized system is shown in Figure 3a. The input decentralized version of the same system is given in Figure 3b. Output centralized and decentralized systems can be represented by directional duals of the digraphs shown in Figure 3.



(a) (b)

Fig. 3. (a) Input Centralized System (b) Input Decentralized System

A convenient way to characterize input and output decentralized systems is to use the interconnection matrix M^* defined in (16). Then a system $\mathcal{S}$ with a condensation $\mathcal{D}^*$ is input decentralized if and only if in M^* of (16), $L^* = I_s$ where I_s is the $s \times s$ identity matrix. Similarly, $\mathcal{S}$ is an output decentralized system if and only if $F^* = I_s$.

By using the methods of reference [15], it is possible to permute the states of a nonlinear system $\mathcal{S}$ and choose an appropriate condensation $\mathcal{D}^*$ of the corresponding digraph $\mathcal{D}$ so that the new transformed system is in either input or output decentralized form or both. In linear systems, however, it is possible to do better [18] and use similarity transformations to produce input and output decentralized systems. Such decentralization procedures are required preliminary steps in stabilization [16], optimization [17], and estimation [18] of linear dynamic systems by decentralized feedback.

By applying Theorem 4 to M^* of (14), we conclude that if $L^* = F^* = I_s$, then the input-output decentralized system $\mathcal{S}$ is input-output reachable whenever each subsystem $\mathcal{S}_i$ is input-output reachable regardless of the form of the state interconnection matrix E^* of $\mathcal{D}^*$. This somewhat trivial result leads to an interesting conclusion that in decentralized systems input-output reachability is invariant under perturbations of the interconnection structure among the subsystems. This aspect of decentralized dynamic systems is expressed precisely by the following corollary to Theorems 6 and 7:

Corollary 2. A system $\mathcal{S}$ is connectively input (output) reachable if it is an input (output) decentralized system.

This corollary announces a strong advantage of decentralized systems over centralized ones, where common inputs are shared by the subsystems. This fact was not stated yet in the open literature on control systems but is probably appreciated intuitively by the authors responsible for a recent fast increase in the number of results in the area of decentralized control [23].

VII. CONCLUSIONS

It has been shown that digraphs can be associated with dynamic systems so that numerous strong results obtained in pure theory of structures can be used to explore various structural properties of dynamic systems. Input and output reachability was formulated as structural counterparts to controllability and observability. It was shown how the reachability properties can be investigated using efficient algorithms developed in digraph theory as applied to computer systems. Then, dynamic systems were considered under the structural perturbations and vulnerability was defined as loss of input and output reachability due to line or point removals from the corresponding digraph. In this context, input and output decentralized systems were identified as invulnerable to arbitrary structural perturbations involving the states of the system.

A large number of problems of reachability and vulnerability were not either mentioned nor discussed in sufficient detail. By restricting structural perturbations to specific interconnections, it should be possible to obtain stronger conditions for invulnerability of dynamic systems. Furthermore, this would open up another possibility of determining the maximum number of lines of the digraph whose removal does not affect reachability of the dynamic systems, as well as the minimum number of lines of the digraph whose presence is required to preserve reachability and the minimum number of lines whose removal destroys it. Solutions to such and other similar problems can be used to further exploit the powerful theory of directed graphs and come up with new and important results in the pure structural analysis of dynamic systems.

REFERENCES

1. Šiljak, D. D., "Connective Stability of Competitive Equilibrium", Automatica, 11(1975), 389-400.
2. Šiljak, D. D., "On Input-Output Reachability of Dynamic Systems", International Journal of Systems Science (to appear).
3. Kalman, R. E., P. L. Falb, and M. A. Arbib, "Topics in Mathematical System Theory", McGraw-Hill, New York, 1969.
4. Harary F., R. Z. Norman, and D. Cartwright, "Structural Models: An Introduction to the Theory of Directed Graphs", Wiley, New York, 1965.
5. Festinger, L., S. Schacter, and K. Back, "Social Pressures in Informal Groups",Harper, New York, 1950.
6. Berztiss, A.T.,"Data Structures: Theory and Practice", Academic Press, New York, 1975.
7. Warshall, S., "A Theorem on Boolean Matrices", Journal of the Association for Computing Machinery, 9(1962), 11-12.
8. Tarjan, R. E., "Depth-First Search and Linear Graph Algorithms", SIAM Journal on Computing, 1(1972), 146-160.
9. Bowie, W. S., "Applications of Graph Theory in Computer Systems", International Journal of Computer and Information Sciences, 5(1976), 9-31.
10. Šiljak, D. D., "When is a Complex Ecosystem Stable?", Mathematical Biosciences, 25(1975), 25-50.
11. Šiljak, D. D., "Competitive Economic Systems: Stability, Decomposition and Aggregation", IEEE Transactions, AC-21(1976), 149-160.
12. Šiljak, D. D., "Stabilization of Large-Scale Systems: A Spinning Flexible Spacecraft", Automatica (to appear).
13. Purdom, P., "A Transitive Closure Algorithm", Nordisk Tidskrift for Informationsbehandling, 10(1970), 76-94.
14. Munro, I., "Efficient Determination of the Transitive Closure of a Directed Graph", Information Proceeding Letters, 1(1971), 56-58.
15. Kevkorian, A. K., "Structural Aspects of Large Dynamic Systems", Proceedings of the Sixth IFAC Congress, Boston, Massachusetts, 19.3(1976), 1-7.
16. Šiljak, D. D., and M. B. Vukčević, "Large Systems: Stability, Complexity, Reliability", Journal of the Franklin Institute, 301(1976), 49-69.

17. Šiljak, D. D., and M. K. Sundareshan, "A multilevel Optimization of Large-Scale Dynamic Systems", IEEE Transactions, AC-21(1976), 79-84.

18. Šiljak, D. D., and M. B. Vukčević, "Decentralization, Stabilization, and Estimation of Large-Scale Linear Systems", IEEE Transactions, AC-21(1976), 363-366.

19. Kalman, R. E., "Mathematical Description of Linear Dynamical Systems", SIAM Journal of Control, 1(1963), 152-192.

20. Lin, C. T., "Structural Controllability", IEEE Transactions, AC-19(1974), 201-208.

21. Shields, R. W., and J. B. Pearson, "Structural Controllability of Multi-input Linear Systems", IEEE Transactions, AC-21(1976), 203-212.

22. Ćalović, M. S., M. V. Djorović, and D. D. Šiljak, "Automatic Generation Control: A Decentralized Approach", The Fourteenth Annual Allerton Conference on Circuit and System Theory, University of Illinois, Monticello, Illinois, Sept. 29-Oct. 1, 1976.

23. Sandel, N. R. Jr., P. Varaiya, and M. Athans, "A Survey of Decentralized Control Methods for Large Scale Systems", Proceedings of the Engineering Foundation Conference on System Engineering for Power: Status and Prospects, Henniker, New Hampshire, 1975, pp. 334-352.

ON STRUCTURAL STABILITY OF MULTIVARIABLE SYSTEMS

E. FREUND

Institut für Informationsverarbeitung in Technik und Biologie
der Fraunhofer-Gesellschaft e.V., 7500 Karlsruhe, West-Germany

Abstract

Various design methods for linear multivariable systems are subject to structural instability if the coefficients of the real physical system have dependencies on state variables and time. This problem is investigated and two conditions for structural stability are presented. As an extension the relations are given for a system representation with an invariant output set.

1. Introduction

A number of state space methods for multivariable systems are based on the derivation of a matrix that characterizes the relation between the inputs of the system and the outputs or derivatives of the outputs uniquely. Linear design methods where this principle occurs are e. g. model matching, system-inversion, input-output decoupling as well as group decoupling, disturbance isolation (e. g. [1] - [5], for other references see [2]). It is well known that this characterizing matrix (in literature often denoted as $\underline{D}^*$) can cause structural instability of the multivariable system, i. e. small variations of the coefficients of the real physical system can result in great changes of the overall behaviour or even in the failure of the whole control system. The subject of the paper is the investigation of this problem and the presentation of two criteria that provide structural stability for variations of the coefficients of the basic system with respect to the state variables and the time.

2. Formulation of the problem

The following system is considered

$$\dot{\underline{x}}(t) = \underline{A}\,\underline{x}(t) + \underline{B}\,\underline{u}(t) \tag{1a}$$

$$\underline{y}(t) = \underline{C}\,\underline{x}(t) + \underline{D}\,\underline{u}(t) \tag{1b}$$

where $\underline{x}(t)$ is the n-dimensional state vector, $\underline{u}(t)$ and $\underline{y}(t)$ are m-dimensional input and output vectors respectively. $\underline{A}$, $\underline{B}$, $\underline{C}$, $\underline{D}$ are of compatible order. The state variable feedback (with $\underline{v}$ as the new m-dimensional input vector and $\underline{F}$ and $\underline{G}$ of compatible order)

$$\underline{u}(t) = \underline{F}\,\underline{x}(t) + \underline{G}\,\underline{v}(t) \tag{2}$$

leads to the closed loop system.

$$\dot{\underline{x}}(t) = (\underline{A} + \underline{B}\,\underline{F})\underline{x}(t) + \underline{B}\,\underline{G}\,\underline{v}(t) \tag{3a}$$

$$\underline{y}(t) = (\underline{C} + \underline{D}\,\underline{F})\underline{x}(t) + \underline{D}\,\underline{G}\,\underline{v}(t) \tag{3b}$$

It is shown in several papers (e. g. [1] - [4]) that system (3) does not have a suitable form for the solution of various control problems like those mentioned in the introduction. The appropriate form for the treatment of these problems is the system representation (4) which was first mentioned in [1] and is based on the concept of the differential order rather than on the structure algorithm like in [2]. Both concepts, however, are in principle transferable though the structure algorithm is somewhat more generalized. For the investigation of structural stability the differential order is more suitable so that this concept is considered here. Then, the system representation has the following form [1]

$$\dot{\underline{x}}(t) = \underline{A}\,\underline{x}(t) + \underline{B}\,\underline{u}(t) \tag{4a}$$

$$\underline{y}^*(t) = \underline{C}^*\underline{x}(t) + \underline{D}^*\underline{u}(t) \tag{4b}$$

where the i-th rows of $\underline{C}^*$ and $\underline{D}^*$ are (i = 1,2,...m)

$$\underline{c}_i^* = \underline{c}_i\,\underline{A}^{d_i} \tag{5}$$

and

$$\underline{D}_i^* = \begin{cases} \underline{D}_i & \text{for } d_i = 0 \\ \underline{c}_i\,\underline{A}^{d_i - 1}\underline{B} & \text{for } d_i \neq 0 \end{cases} \tag{6}$$

with $\underline{c}_i$ as i-th row of $\underline{C}$. $\underline{y}^*$ is given by

$$\underline{y}^* = [\,y^{(d_1)} \;\ldots\; y^{(d_m)}\,]' \tag{7}$$

and d_i is the differential order defined by

$$d_i = \min\{\, j : \underline{c}_i\,\underline{A}^{j-1}\underline{B} \neq 0\,;\; j = 1,2,\cdots m \,\} \tag{8}$$

Then, the closed loop system (3) has the useful form

$$\dot{\underline{x}}(t) = (\underline{A} + \underline{B}\,\underline{F})\underline{x}(t) + \underline{B}\,\underline{G}\,\underline{v}(t) \tag{9a}$$

$$\underline{y}^*(t) = (\underline{C}^* + \underline{D}^*\underline{F})\underline{x}(t) + \underline{D}^*\underline{G}\,\underline{v}(t) \tag{9b}$$

where $\underline{D}^*$ is assumed to have full rank.

As examples for various control problems based on system (9), decoupling and model matching are mentioned. It was shown in [1] that system (3) is input-output decoupled (i. e. $v_i(t)$ effects only $y_i(t)$ for i = 1,2,...m) if

$$\underline{F} = -\underline{D}^{*-1}(\underline{C}^{*}+\underline{M}) \quad ; \quad \underline{G} = \underline{D}^{*-1}\underline{\Lambda} \tag{10}$$

where the i-th rows of $\underline{M}$ and the diagonal matrix Λ are

$$\underline{M} = \begin{cases} 0 & \text{for } d_i = 0 \\ \sum\limits_{k=0}^{d_i-1} \alpha_k^i \underline{c}_i \underline{A}^k & \text{for } d_i \neq 0 \end{cases} \quad ; \underline{\Lambda}_i = [0 \cdots 0 \lambda_i 0 \cdots 0] \tag{11}$$

In model matching $\underline{F}$ and $\underline{G}$ in the feedback equation (2) have to be chosen such that the closed loop system has a desired overall transition matrix H_R. It was shown in [3] that instead of

$$\underline{H}_R = \underline{H}(\underline{A}+\underline{B}\underline{F}, \underline{B}\underline{G}, \underline{C}+\underline{D}\underline{F}, \underline{D}\underline{G})$$

which is related to (3), the relation

$$\underline{H}_R = \underline{H}(\underline{A}+\underline{B}\underline{F}, \underline{B}\underline{G}, \underline{C}^{*}+\underline{D}^{*}\underline{F}, \underline{D}^{*}\underline{G}) \tag{12}$$

corresponding to (9) is basic for the solution of the model matching problem. The solutions to the other problems mentioned like system inversion, group decoupling, disturbance isolution etc. are also based on system representation (9).

It is obvious from the presented equation (9) that the matrix $\underline{D}^*$ (characterizing the input-output relation of system (1)) must maintain full rank and that the differential order d_i ($i = 1,2,\ldots m$) must remain constant for all states and time in the range of the actual application. This is of course the case for the consideration of a physical system as a system with constant coefficients like in (1). Real physical system, however, have often nonlinearities or variations of the coefficients with respect to the states and the time. This can lead to structural instabilities of the multivariable system due to changes in the differential order or in the rank of the characterizing matrix $\underline{D}^*$. Hereby, structural instability means that small parameter variations can lead to great changes of the overall systems behaviour. In the example of decoupling, this can easily be seen from equations (6) and (8) in connection with (10) and (11). A change of any of the differential orders d_i or of the rank of $\underline{D}^*$ due to small variations of parameters would lead to an unappropriate feedback and feedforward path (10) with great changes of the system behaviour.

3. Conditions for structural stability

In real physical systems, at least some of the coefficients of the matrices $\underline{A}$, $\underline{B}$, $\underline{C}$, $\underline{D}$ of system (1) have dependencies on the state variables and the time so that a more realistic system representation than (1) would be

$$\dot{\underline{x}}(t) = \underline{\tilde{A}}(\underline{x},t) + \underline{\tilde{B}}(\underline{x},t)\,\underline{u}(t) \tag{13a}$$

$$\underline{y}(t) = \underline{\tilde{C}}(\underline{x},t) + \underline{\tilde{D}}(\underline{x},t)\,\underline{u}(t) \tag{13b}$$

In (13) $\underline{A}(\underline{x},t)$, $\underline{B}(\underline{x},t)$, $\underline{C}(\underline{x},t)$ and $\underline{D}(\underline{x},t)$ are n x 1, n x m, m x 1 and n x m matrices, respectively. The question is whether the design of the overall system based on system representation (1) is structurally stable in consideration of (13). For the design problems described before the following result can be given as solution to this problem:

The use of system representation (1) leads to a structurally stable system inspite of nonlinearities as presented in (13) if for all x and t in the considered domain the following two conditions are both satisfied:

Condition 1

The differential order defined by

$$d_i = \min\left\{ j : \left[\frac{\partial}{\partial \underline{x}} N_{\tilde{A}}^{j-1}\tilde{C}_i(\underline{x},t)\right]\underline{\tilde{B}}(\underline{x},t) \neq \underline{0},\ j=1,2,\dots m\right\} \tag{14}$$

remains constant for all outputs Y_1, Y_2, ... Y_m (which is i = 1,2,...m)

and

Condition 2

the rank of the matrix $\underline{\tilde{D}}^*(\underline{x},t)$ with the rows (i = 1,2,...m)

$$\underline{\tilde{D}}_i^*(\underline{x},t) = \begin{cases} \underline{\tilde{D}}_i(\underline{x},t) & \text{for } d_i = 0 \\ \left[\frac{\partial}{\partial \underline{x}} N_{\tilde{A}}^{d_i-1}\tilde{C}_i(\underline{x},t)\right]\underline{\tilde{B}}(\underline{x},t) & \text{for } d_i \neq 0 \end{cases} \tag{15}$$

remains constant.

In (14) and (15), $\underline{\tilde{D}}_i(\underline{x},t)$ is the i-th row of $\underline{\tilde{D}}(x,t)$ and $\tilde{C}_i(x,t)$ is the i-th component of $\underline{\tilde{C}}(\underline{x},t)$ for i = 1,2,...m; the nonlinear operator is defined by

$$N_{\tilde{A}}^{k}\tilde{C}_i(\underline{x},t) = \frac{\partial}{\partial t}N_{\tilde{A}}^{k-1}\tilde{C}_i(\underline{x},t) + \left\{\frac{\partial}{\partial \underline{x}} N_{\tilde{A}}^{k-1}\tilde{C}_i(\underline{x},t)\right\}\underline{A}(\underline{x},t) \tag{16}$$

for k = 1,2,... with $N_A^0\,\tilde{C}_i(\underline{x},t) = \tilde{C}_i(\underline{x},t)$

For the proof of conditions 1 and 2 with eqs. (14) and (15) the i-th output of system (13), $Y_i(t)$, is considered [6]:

$$y_i(t) = \tilde{C}_i(\underline{x},t) + \underline{\tilde{D}}_i(\underline{x},t)\cdot\underline{u}(t) \tag{17}$$

For $d_i > 0$, $\underline{\tilde{D}}_i$ equals zero and $Y_i(t)$ is differentiated with respect to t

$$\dot y_i(t) = \frac{\partial \tilde C_i(\underline x,t)}{\partial t} + \frac{\partial \tilde C_i(\underline x,t)}{\partial \underline x}\cdot\frac{\partial \underline x}{\partial t}$$

Substitution of (13a) and application of the operator (16) leads to

$$\dot y_i(t) = N_{\tilde A}\tilde C_i(\underline x,t) + \left[\frac{\partial \tilde C_i(\underline x,t)}{\partial \underline x}\right]\underline{\tilde B}(\underline x,t)\,\underline u(t)$$

which simplifies to

$$\dot y_i(t) = N_{\tilde A}\tilde C_i(\underline x,t)$$

if the differential order defined in (14) is $d_i > 1$. Repeatition of this procedure gives for the differential order $d_i > 0$ the general equation

$$y_i^{(d_i)}(t) = N_{\tilde A}^{d_i}\tilde C_i(\underline x,t) + \left[\frac{\partial}{\partial \underline x} N_{\tilde A}^{d_i-1}\tilde C_i(\underline x,t)\right]\underline{\tilde B}(\underline x,t)\,\underline u(t) \tag{18}$$

For all outputs (i = 1,2,...m) (eq. (7)) follows, therefore

$$\underline{\dot x}(t) = \underline{\tilde A}(\underline x,t) + \underline{\tilde B}(\underline x,t)\,\underline u(t) \tag{19a}$$

$$\underline y^*(t) = \underline{\tilde C}^*(\underline x,t) + \underline{\tilde D}^*(\underline x,t)\,\underline u(t) \tag{19b}$$

with $\underline{\tilde D}^*$ as in equation (15) and

$$\tilde C_i^*(\underline x,t) = N_{\tilde A}^{d_i}\tilde C_i(\underline x,t) \tag{20}$$

as i-th component of $\underline{\tilde C}^*$ (i = 1,2,...m)

The nonlinear system (19) is related to the linear system (4) (corresponding to systems (1) and (13), respectively). From this follows that the linear system representation (4) and (9) are only structurally stable if the conditions 1 and 2 in (14) and (15) are fulfilled.

4. Extension to the partitioned system

If system (1) has a different number of inputs and outputs or if D* in (4b) is not of full rank the solution to the group of control problems that were mentioned before requires a partitioned system representation [2, 3, 5]. This system has an additional output set that is independent of the inputs of the system for all derivation of the outputs with respect to time. This means that in these cases the investigation of structural stability requires a partitioned system representation of the type

$$\underline{\dot x}(t) = \underline{\tilde A}(\underline x,t) + \underline{\tilde B}(\underline x,t)\,\underline u(t) \tag{21a}$$

$$\underline y_a^*(t) = \underline C_a^*(\underline x,t) + \underline D_a^*(\underline x,t)\,\underline u(t) \tag{21b}$$

$$\underline y_b^*(t) = \underline C_b^*(\underline x,t) \tag{21c}$$

which corresponds to the nonlinear form (13). Form (21) can be derived like system representation (19) in principle. The determination of the maximal rank of $\underline{D}^*_a$ and of the output set $\underline{Y}^*_b$, however, can cause some difficulties. Form (21) is constructed such that only $\underline{Y}^*_a$ and $\underline{u}$ with $\underline{D}^*_a$ characterize the input-output relation while $\underline{Y}^*_b$ is independent of the inputs of the system and depends only on the intial state $\underline{x}_o$. For the derivation of the partitioned system the following steps can be used:

a) The procedure for the derivation of (18) together with (14) and (15) is applied to all outputs,

b) the linearly independent rows $\underline{\tilde{D}}_i(\underline{x},t)$ $(i = 1,2,...q)$ are gathered in $\underline{D}^*$ (eventually by reordering the outputs). It is checked by additional differentiation steps on the remaining inputs like in a) whether the rank of $\underline{D}^*_a$ can be increased or if it is the maximal rank. In practical cases the rank of $\underline{D}^*_a$ is often directly terminated by the number of inputs.

c) For remaining outputs, an eventual direct dependency on the inputs is checked using the derivation procedure of (18) for n differentiation steps. If a direct dependency occurs a nonlinear transformation is used to transform this term into $\underline{D}^*_a$.

This principal derivation of form (21) shows that the structural stability of the system with constant coefficients that corresponds to system (21) is guaranteed if conditions (14) and (15) are fulfilled. The satisfaction of these conditions is especially important for the output set $\underline{Y}^*_a$ in (21b) because a violation of the conditions would lead to a change of the overall systems behaviour which was described before. A violation of these conditions for the output set $\underline{Y}^*_b$, however, does not cause structural instability because this set does not influence the overall behaviour. $\underline{Y}^*_b$, however, would lose to some extend its feature of input invariance. This means that for example in group decoupling a dependency of the outputs $\underline{Y}^*_b$ on the corresponding input set can occur while the overall group decoupling behaviour remains unchanged.

5. Examples

Example 1

To demonstrate condition 1 for structural stability, the following simple example is considered:

$$\overset{\circ}{\underline{x}} = \begin{bmatrix} 0 & 1 & 2 \\ -1 & 0 & 2 \\ 1 & 1 & 0 \end{bmatrix} \underline{x} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{22}$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1+f_1(x_1,x_3) & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \underline{x}$$

Deleting the nonlinearity $f_1(x_1,x_3)$, gives from (8) for the first output variable y_1:

$$\underline{C}_1 \underline{B} = [1 \quad 0] \neq \underline{0} \tag{23}$$

which is $d_1 = 1$. If the dependency of the coefficient on the states x_1 and x_3 is taken into account, the application of (14) leads to

$$\left[\frac{\partial}{\partial \underline{x}} \tilde{C}_1(\underline{x},t)\right] \tilde{\underline{B}}(\underline{x},t) = \left[x_1 \left\{ \frac{\partial f_1(x_1,x_3)}{\partial x_1} + \frac{\partial f_1(x_1,x_3)}{\partial x_3} \right\} + f_1(x_1,x_2) + 1 \;\vdots\; 0 \right] \tag{24}$$

This means from condition 1 (eq. 14) that system (22) is only structurally stable for deleting $f_1(x_1,x_3)$ if for all x_1 and x_3

$$x_1 \left\{ \frac{\partial f_1(x_1,x_3)}{\partial x_1} + \frac{\partial f_1(x_1,x_3)}{\partial x_3} \right\} + f_1(x_1,x_3) + 1 \neq 0 \tag{25}$$

or otherwise the differential order d_1 will change. Condition (25) shows that $f_1(x_1,x_3) \ll 1$ matters only to some extend for structural stability though it looks obvious from (22). If structural stability in the decoupling problem is considered, it follows from (9), (10) and (11) for the overall behaviour of the decoupled system (22) in case $f_1(x_1,x_3)$ is deleted

$$\begin{aligned} \overset{\circ}{y}_1 + \alpha_0^1 y_1 &= \lambda_1 v_1 \\ \overset{\circ}{y}_2 + \alpha_0^2 y_2 &= \lambda_2 v_2 \end{aligned} \tag{26}$$

However, if the left side of (25) equals zero, i. e. if condition 1 is not satisfied, the equation for y_1 is structurally unstable:

$$\begin{aligned} \overset{\circ}{y}_1 &= y_2 - 2x_3 \\ \overset{\circ}{y}_2 + \alpha_0^2 y_2 &= \lambda_2 v_2 \end{aligned} \tag{27}$$

Example 2

To demonstrate condition 2, the following example is given:

$$\overset{\circ}{\underline{x}} = \begin{bmatrix} 1 & 0 & 1 \\ -1+f_2(x_1,x_2) & 0 & 2 \\ 0 & 1 & 0 \end{bmatrix} \underline{x} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{28}$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \underline{x}$$

If $f_2(x_1,x_2)$ is deleted with respect to -1, it follows from (8) and (6) $d_1 = 1$, $d_2 = 2$ and

$$\underline{D}^* = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \tag{29}$$

i. e. rank $\underline{D}^* = 2$, so that the system can - for example - be decoupled (eqs. (10) and (11)). When the parameter variation $f_2(x_1,x_2)$ is taken into account, however, it follows from (14) and (15)

$$\underline{\tilde{D}}^* = \begin{bmatrix} 1 & 1 \\ 1 + x_1 \dfrac{\partial f_2(x_1,x_2)}{\partial x_1} + f_2(x_1,x_2) & 2 \end{bmatrix} \tag{30}$$

The rank of $\underline{\tilde{D}}^*$ remains only constant, i. e. system (28) is structurally stable from condition 2 in spite of deleting $f_2(x_1,x_2)$ if for all x_1 and x_2

$$f^*(x_1,x_2) \neq 1 \tag{31}$$

where

$$f^*(x_1,x_2) = x_1 \cdot \frac{\partial f_2(x_1,x_2)}{\partial x_1} + f_2(x_1,x_2) \tag{32}$$

If as an application structural stability of the decoupled system is considered, it follows from (9), (10) and (11) in connection with (29) and (30)

$$\begin{aligned} &\ddot{y}_1 + \alpha_1^1 \dot{y}_1 + \alpha_0^1 y_1 = \lambda_1 V_1 \\ &\ddot{y}_2 + (1-f^*(x_1,x_2))\alpha_1^2 \dot{y}_2 + (1-f^*(x_1,x_2))\alpha_0^2 y_2 \\ &\qquad = (1-f^*(x_1,x_2))\lambda_2 V_2 + f^*(x_1,x_2)(\alpha_0^1 y_1 + \lambda_1 V_1) \end{aligned} \tag{33}$$

Eqs. (33) demonstrate what happens if condition 2 is not satisfied: The equation for Y_2 is only driven by Y_1 and V_1, i. e. in the overall behaviour Y_2 is not decoupled anymore. It follows from (32) that the case $f^*(x_1,x_2) = 0$ describes the linear system with constant coefficients with the desired overall behaviour. The closer $f^*(x_1,x_2)$ gets to 1, the more the dynamics of the system are changed and the stronger are the couplings. This trend is not specific for this example but follows directly from condition 2 (eq. (15) for structural stability so that (33) does not have to be to determined in practical applications.

References

[1] P.L. Falb and W.A. Wolovich, "On the decoupling of multivariable control systems", IEEE Trans. Automatic Control, AC-12 (1967), pp. 651-6.

[2] L.M. Silverman and H.J. Payne, "Input-output structure of linear systems with application to the decoupling problem", SIAM J. on Control, vol. 9 (1971), pp. 199-233.

[3] B.C. Moore and L.M. Silverman, "Model matching by state feedback and dynamic compensation", IEEE Trans. AC-17 (1972), pp. 491-497.

[4] E. Freund, "Zeitvariable Mehrgrößensysteme", Springer Verlag, Berlin-Heidelberg, New York 1971.

[5] S.M. Sato and P.V. Lopresti, "On the generalization of state feedback decoupling theory", IEEE Trans. Automatic Control, AC-16 1971), pp. 133-139.

[6] E. Freund, "The structure of decoupled nonlinear systems", Int. J. Control, vol. 21 (1975), pp. 443-450

ON A RESULT BY HESTENES AND ITS APPLICATION IN BILINEAR SYSTEMS STABILIZATION

TUDOR IONESCU
Department of Guidance, Navigation and Control
Intermetrics, Inc., Cambridge, Mass. 02138

ABSTRACT

Sufficient conditions for the existence of constant stabilizing controls for several classes of continuous bilinear systems are obtained via a theorem concerning quadratic forms due to Hestenes [1]. It is shown that the set U from which such controls can be selected is convex and open. A computational procedure that permits complete determination of U via standard eigenvalue methods is also given.

INTRODUCTION

Let E be a linear subspace of R^n. Let R(x) and Q(x) be two quadratic real forms on E, in the sense that there exist real-valued bilinear forms R(x,y), Q(x,y) on E X E such that R(x) = R(x,x) and Q(x) = Q(x,x). Moreover assume that R and Q are independent on E, i.e., there exist no multipliers λ, μ, (not both zero) such that: $\lambda R(x) + \mu Q(x) = 0$ on E.

The following theorem is due to Hestenes [1]:

Theorem H: Suppose that R(x) = 0, Q(x) = 0 only in case x = 0, and that the image W = (u,v) of E under the transformation u = R(x), v = Q(x) is closed. If the dimension of E is two, suppose further that W is not the uv plane. Then the set of multipliers (λ,μ) given by:

$$U^{-}(R,Q) \underline{\underline{\Delta}} \{(\lambda,\mu) \mid \lambda R(x) + \mu Q(x) < 0, \ \forall x \varepsilon E\}$$

is the interior of U(R,Q):

$$U\ (R,Q) \underline{\underline{\Delta}} \{(\lambda,\mu) \mid \lambda R(x) + \mu Q(x) \leq 0, \ \forall x \varepsilon E\}$$

Moreover, U possesses two independent generators (in the sense defined in [1]).

It is intuitively clear that results of this type can be used in solving standard Lyapunov stability problems. In particular, stabilization properties for the class of bilinear systems described by

$$\dot{x} = (\sum_{i=1}^{m} u_i(t) \cdot A_i)x \qquad (I.1)$$

where $A_i \varepsilon R^{nxn}$, $u_i \varepsilon R$ appears to be quite naturally associated with results such as Theorem H. The main purpose of this paper is to investigate this association by discussing constant control stabilization for (I.1). For the bilinear systems with drift term (i.e., $u_1 \equiv 1$), one can actually construct a set of stabilizing controls via a computational procedure no more complex than the usual matrix eigenvalue-eigenvector calculation.

The nonconstant control stabilization for (I.1) has been first discussed in [2]. This paper includes a straightforward generalization of the class of stabilizing feedback controls as given by Proposition 1 in [2]. We conclude with a few simple numerical examples.

1. STABILIZATION BY CONSTANT CONTROLS

Let us write again the bilinear system equation:

$$\dot{x}(t) = \left(\sum_{i=1}^{m} u_i(t) A_i \right) x(t) \tag{1.1}$$

where $A_i \in R^{n \times n}$ are linearly independent, $u_i(t)$ being piecewise continuous functions. As previously mentioned, in this section we are interested in finding sufficient conditions for the existence of constant controls u_i that will stabilize $x \equiv 0$. Note the trivial fact that $x \equiv 0$ is a stable equilibrium point of (1.1) if the controls are identically zero.

Before proceeding with the first theorem certain definitions are in order. Let $x^T M x$ and $x^T N x$ be two real valued quadratic forms on R^n, $M, N \in R^{n \times n}$ linearly independent. Then we denote by $U(M,N)$, $U^+(M,N)$, $U^-(M,N)$ the following sets:

$$U(M,N) \triangleq \{(\lambda,\mu) \mid \lambda \cdot x^T M x + \mu \cdot x^T N x \leq 0, \ \forall x \in R^n\},$$

$$U^+(M,N) \triangleq \{(\lambda,\mu) \mid \lambda \cdot x^T M x + \mu \cdot x^T N x > 0, \ \forall x \in R^n\},$$

$$U^-(M,N) \triangleq \{(\lambda,\mu) \mid \lambda \cdot x^T M x + \mu \cdot x^T N x < 0, \ \forall x \in R^n\}$$

Also for some symmetric matrix $P \in R^{n \times n}$ denote: $Q_i \triangleq PA_i + A_i^T P$, $i = 1,\dots,m$. Now:

Theorem (1.1)

Suppose that there exist a matrix $P = P^T > 0$ and two integers $j_i, j_2 \in \{1,\dots,m\}$ such that

$$\{x \in R^n \mid x^T Q_{j_i} x = 0\} \cap \{x \in R^n \mid x^T Q_{j_2} x = 0\} = \{0\} \tag{1.2}$$

If $n = 2$, suppose further that there is a point $(\alpha,\beta) \in R^2$ that is not of the form

$$\alpha = x^T Q_{j_1} x \quad , \quad \beta = x^T Q_{j_2} x \tag{1.3}$$

for some $x \in R^2$. Then the system (1.1) is stabilized by the constant controls: $u_i = 0$, $i \neq j_1, j_2$; $u_{j_1} = \lambda$, $u_{j_2} = \mu$ for $(\lambda,\mu) \in U^-(Q_{j_1}, Q_{j_2})$.

Proof. Let $V(x) = x^T P x$ be a Lyapunov function candidate for (1.1). Clearly $x^T P x > 0$ for all $x \neq 0$. Then:

$$\dot{V}(x(t)) = \sum_{i=1}^{m} u_i \, x^T (A_i^T P + PA_i) x \tag{1.4}$$

It suffices to show that $V(x) < 0$ for $x \neq 0$ for some choice of $\underline{u}_i$. Let W be the set: $W = \{(\alpha,\beta) \mid \alpha = x^T Q_{j_1} x,\ \beta = x^T Q_{j_2} x,\ x \varepsilon R^n\}$. Hestenes [1] shows that W is a convex cone with the property that if $(\alpha,\beta)\varepsilon W$, then $(a\alpha, a\beta)\varepsilon W$ for all $a \leq 0$.

Under these assumptions, it follows that W is also closed. Indeed, basically repeating the procedure given in [1], let (α_o, β_o) be a limit point of W that is not in W. One can choose a sequence of vectors $\{x_k\} \varepsilon R^n$ such that

$$\lim_{k\to\infty} x_k^T Q_{j_1} x_k = \alpha_o, \quad \lim_{k\to\infty} x_k^T Q_{j_2} x_k = \beta_o \tag{1.5}$$

If $|x_k|$ is the absolute value norm, it follows that $\lim_{k\to\infty} |x_k| = +\infty$, since otherwise $\{x_k\}$ would converge to a point $x_o \varepsilon R^n$ such that

$$x_o^T Q_{j_1} x_o = \alpha_o, \quad x_o^T Q_{j_2} x_o = \beta_o \tag{1.6}$$

which is impossible since $(\alpha_o, \beta_o) \notin W$. Now replace $\{x_k\}$ by a subsequence such that the unit vectors $z_k = \frac{x_k}{|x_k|}$ converge to a vector $z_o \neq 0$. We have:

$$z_o^T Q_{j_1} z_o = \lim_{k\to\infty} z_k^T Q_{j_1} z_k = \lim_{k\to\infty} \frac{x_k^T Q_{j_2} x_k}{|x_k|^2} = 0$$

and

$$z_o^T Q_{j_2} z_o = \lim z_k^T Q_{j_2} z_k = \lim \frac{x_k^T Q_{j_2} x_k}{|x_k|^2} = 0 \tag{1.7}$$

which is impossible. It follows that W is closed.

One can now apply theorem H to $x^T Q_{j_1} x$ and $x^T Q_{j_2} x$. Indeed, this is possible since by assumption $x^T Q_{j_i} x = 0$ and $x^T Q_{j_1} x = 0$ only if $x=0$. Moreover, we showed that W is closed. Therefore, there exist (λ_i, μ_i) $i = 1,2$ such that:

$$\lambda_i \cdot x^T Q_{j_1} x + \mu_i \cdot x^T Q_{j_2} x < 0 \tag{1.8}$$

i.e., $(\lambda_i, \mu_i) \ \varepsilon\ U^-(Q_{j_1}, Q_{j_2})$

Thus by choosing the controls u_i, $i\varepsilon\{1,\ldots,m\}$ as suggested, we can conclude that $\dot{V}(x) < 0$. Also, one should notice that if (λ_1, μ_1) and (λ_2, μ_2) belong to U (or U^-), i.e.:

$$\lambda_1 \cdot R(x) + \mu_i \cdot Q(x) \leq 0 \qquad i = 1,2 \tag{1.9}$$

then (λ_i,μ_i) will generate U(or U^-) in the sense that: $\lambda = a\cdot\lambda_1 + b\cdot\lambda_2$ and $\mu = a\cdot\mu_1 + b\cdot\mu_2$ ($a \geq 0$, $b \geq 0$, $a+b > 0$) belong also to U(or U^-).

Q.E.D.

Remark. The additional requirement concerning the case $n=2$ is necessary to ensure that the closure $\overline{W}$ of W does not contain a half-plane, i.e., it consists of all the points (α,β) satisfying a set of independent inequalities of the form:

$$\lambda_1\cdot\alpha + \mu_1\cdot\beta \leq 0, \qquad \lambda_2\cdot\alpha + \mu_2\beta \leq 0. \tag{1.10}$$

For $n>2$, this is ensured by the first assumption of the theorem.

Theorem (1.1) gives an interesting description for the set of stabilizing controls. In general for n finite U cannot have less than two generators, and U^- is either empty or contains the interior of U. Moreover, U^- is open (see also [1]). The next theorem will show that for bilinear systems with a drift term:

$$\dot{x} = (A + \sum_{i=1}^{m} u_i \cdot A_i)x \tag{1.11}$$

global asymptotic stabilization can be achieved in similar fashion. First, let us denote $Q \triangleq AP + PA^T$ for some symmetric matrix P.

Theorem (1.2)

If there exist a positive definite matrix P and an integer $j_1 \varepsilon \{1,\ldots,m\}$ such that

$$\{x\varepsilon R^n | x^T Q_{j_1} x = 0\} \subseteq \{x\varepsilon R^n | x^T Q x < 0\} \tag{1.12}$$

then the system (1.11) is globally stabilizable by constant controls.

Proof. Following the same procedure as in Theorem (1.1), one can write:

$$\dot{V}(x) = x^T Q x + \sum_{i=1}^{m} u_i\, x^T Q_i\, x$$

Since the assumptions of Theorem (1.1) are met, the set $U^-(Q,Q_{j_1})$ is open and non-empty. Now, if $x^T Q_{j_1} x \neq 0$ whenever $x \neq 0$, then $x^T Q_{j_1} x$ or $-x^T Q_{j_1} x$ is negative on R^n, and there are multipliers of the form $\lambda=0$, $\mu\neq0$ in $U^-(Q,Q_{j_1})$. But since U^- is open, one can modify the multipliers so that $\lambda>0$. These multipliers, in turn, can be scaled so that $\lambda=1$. If $x^T Q_{j_1} x = 0$ for some $x \neq 0$, then $\lambda>0$ for every set $(\lambda,\mu)\varepsilon U^-(Q,Q_{j_1})$, and one can scale again to have $\lambda=1$. Thus, there exists the real number such that the pair $(1,\mu)\varepsilon U^-(Q,Q_{j_1})$. Choosing $u_i=0$ $i\neq j_1$; $u_{j_1}=\mu$ for $(1,\mu)\varepsilon U^-$, one has $\dot{V}(x) < 0$.

Q.E.D.

For the system (1.11) with the assumption (1.12), the set U^- can be completely and uniquely determined in terms of the eigenvalues of the matrix $A + \nu \cdot A_{j_1}$, $\nu \in R$. To show this, let: $D(\nu) = |A + \nu \cdot A_{j_1}|$ denoting the determinant of the pencil.

If A_{j_1} has rank n-r, $0 \leq r < n$, then it is well-known that $D(\nu)$ is a polynomial of degree n-r in ν([4] p. 438). Let:

$$\nu_1 \leq \nu_2 \leq \ldots \leq \nu_{n-r} \tag{1.13}$$

be its zeros, each repeated a number of times equal to its multiplicity. With every root ν of $D(\nu)$, one can associate the number ℓ of linearly-independent solutions of the equation

$$(A + \nu \cdot A_{j_i})\, y_s = 0 \qquad s = 1,\ldots,\ell \tag{1.14}$$

ℓ is called the index of ν.

It is well known that, under the above assumptions, the zeros of $D(\nu)$ are real, and the corresponding solutions of (1.14) may be chosen real. Moreover, for every root, its index equals its multiplicity ([4] p. 430). Thus, for every set of zeros (1.13), one can compute the corresponding linearly-independent vectors $y_1,\ldots,y_{n-r}$ verifying (1.14).

We now state the following proposition which could be regarded as a corollary of theorem (1.2).

Proposition (1.1)

Under the assumption (1.12), the set of stabilizing controls for (1.11) is uniquely determined by:

$$U^-(Q,Q_{j_1}) = \{(1,\mu)\} \text{ where: } \begin{cases} \mu < \nu_1 \text{ if } y_1^T Q_{j_i} y_1 > 0; \\ \mu < \nu_{n-r} \text{ if } y_{n-r}^T Q_{j_1} y_{n-r} < 0; \\ \nu_s < \mu < \nu_{s+1} \text{ if } \begin{cases} y_s^T Q_{j_1} y_s < 0 \\ \text{and} \\ y_{s+1}^T \cdot Q_{j_1} \cdot y_{s+1} > 0 \end{cases} \end{cases}$$

Proof. The proof is based on a series of lemmas due to Reid and proven in [4]:

Lemma 1.1. If ν_1 and ν_2 are two distinct zeros of $D(\nu)$, and y_1, y_2 are the corresponding solutions of (1.14), then:

$$y_1^T Q y_2 = y_1^T Q_{j_1} y_2 = 0$$

Proof. (See lemma 2, [4], p. 438)

Lemma 1.2. If z_k, k=1,...,r are the solutions of the equation $A_{j_1} z = 0$, then: (i) $\{z_k, k=1,\ldots,r\}$; $\{y_s, s=1,\ldots,n-r\}$ are linearly independent (thus they form a basis for R^n); (ii) if $x = \sum_{1}^{r} \alpha_k z_k + \sum_{s=1}^{n-r} \beta_s \cdot y_s$ is some vector in R^n then: $(ii)_a$ $y_s^T \cdot Q_{j_1} \cdot y_s \neq 0$; $y_s^T \cdot Q_{j_1} y_t = 0$, for $\nu_s \neq \nu_t$, $t \varepsilon \{1,\ldots,n-r\}$; $(ii)_b$ $\sum_{k,\ell=1}^{r} \alpha_k \cdot \alpha_\ell \cdot z_k^T Q z_\ell \leq 0$; $(ii)_c$ $z_k^T \cdot Q_z y_s = 0$, $z_k^T \cdot Q_{j_1} y_s = 0$, $k=1,\ldots,r$; $s=1,\ldots,n-r$.

Proof. (See [4] lemma 5 and its corollary, p. 440). The next result is a slight extension of lemma 6 in [4].

Lemma 1.3. Let s be some integer in $\{1,\ldots,n-r\}$

(a) If $y_s^T \cdot Q_{j_1} y_s < 0$, then, for each t such that $\nu_s \geq \nu_t$, one also has

$$y_t^T \cdot Q_{j_1} y_t < 0$$

(b) If $y_s^T \cdot Q_{j_1} y_s > 0$, then, for each t such that $\nu_t \geq \nu_s$, one has

$$y_t^T \cdot Q_{j_1} y_s > 0$$

Proof. By assumptions

$$(A + \nu_s \cdot A_{j_1}) \cdot y_s = 0, \quad s = 1,\ldots,n-r \tag{1.15}$$

Multiplying (1.15) to the left by $y_s^T P$ and taking the transpose, one obtains:

$$y_s^T Q y_s + \nu_s \cdot y_s^T Q_{j_1} y_s = 0 \quad s = \{1,\ldots,n-r\} \tag{1.16}$$

Now, for case (a), suppose that $\nu_s \geq \nu_t$ but $y_t^T Q_{j_1} y_t > 0$. Let

$$c^2 = -\frac{y_t^T Q_{j_1} y_t}{y_s^T Q_{j_1} y_s}$$ (the denominator is different from zero (lemma 1.2 $(ii)_a$)).

Then:

$$(y_t^T + c \cdot y_s^T) Q_{j_1} (y_t + c \cdot y_s) = y_t^T Q_{j_1} y_t + c^2 \cdot y_s^T Q_{j_1} y_s = 0 \tag{1.17}$$

Similarly, using (1.16):

$$\begin{aligned}(y_t^T + c \cdot y_s^T) Q (y_t + c \cdot y_s) &= y_t^T Q y_t + c^2 \cdot y_s^T Q y_s = \\ &= (\nu_s - \nu_t) y_t^T Q_{j_1} y_t \geq 0\end{aligned} \tag{1.18}$$

But (1.17) and (1.18) contradict (1.12), thus concluding the proof of case (a). For case (b), one proceeds in an identical fashion. Now $\nu_s - \nu_t \leq 0$ (1.18), remaining nonnegative, again contradicting the hypothesis.

Q.E.D.

We now return to the proof of our proposition. Let Γ denote the interval of ν on the real axis defined as suggested by the proposition statement. It is clear that Γ is uniquely determined by the Lemma (1.3). Moreover, $(\nu-\nu_s)\ y_s^T\ Q_{j_1}\ y_s$ is negative for every $s = \{1,\dots,n-r\}$, whenever $\nu\varepsilon\Gamma$.

We now claim that the set $\{(1,\mu),\ \mu\varepsilon\Gamma\}$ is in fact $U^-(Q,Q_{j_1})$. Indeed, it is enough to show that for any $\nu\varepsilon\Gamma$, $x^T\ Qx + \nu\cdot x^T\ Q_{j_1}\ x < 0$ for $\nabla x\varepsilon R^n$.

But according to our choice of basis in R^n, $x = \sum_{k=1}^{r} \alpha_k\cdot z_k + \sum_{s=1}^{n-r} \beta_s\cdot y_s$ for some α_k, $\beta_s\ \varepsilon\ R$, not all zero. Therefore using Lemmas (1.1) and (1.2), one has:

$$x^T Q\ x + \nu\cdot x^T\ Q_{j_1}\ x = \sum_{k,\ell=1}^{r} \alpha_k\cdot\alpha_\ell + z_k^T\cdot Qz_\ell + \sum_{s=1}^{n-r} \beta_s^2\cdot y_s^T\ Q\ y_s + \nu\cdot \sum_{s=1}^{n-r} \beta_s^2\cdot y_s^T\ Q_{j_1}\ y_s \tag{1.19}$$

Replacing (1.16) in (1.19) we get:

$$x^T\ Qx + \nu\cdot x^T\ Q_{j_1}\ x = \sum_{\ell=1}^{r} \alpha_k\cdot\alpha_\ell\cdot z_k^T\ Qz_\ell + \sum_{s=1}^{n-r} (\nu-\nu_s)\cdot y_s^T\ Q_{j_1}\ y_s \tag{1.20}$$

From Lemma 1.2 ($(ii)_b$) and the fact that $(\nu-\nu_s)\cdot y_s^T\ Q_{j_1}\ y_s$ is negative for every $s\varepsilon\{1,\dots,n-r\}$ whenever $\nu\varepsilon\Gamma$, it follows that the R.H.S. of 1.20 is negative definite for $\nabla x\varepsilon R^n$, and we are done.

Q.E.D.

Example

Let $\dot{x} = \left(\begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix} + u \begin{bmatrix} 0 & 1 \\ 1 & c_1^2 \end{bmatrix}\right) x$. Choosing $P = \frac{1}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

one has $x^T\ Q\ x = -\ x_1^2$, $x^T\ Q_1\ x = 2x_1\ x_2 + c_1^2\ x_2^2$.

Clearly, $x^T\ Q\ x < 0$ for $x^T\ Q_1\ x = 0$, thus theorem 1.2 applies.

In order to determine the stabilizing controls, one can proceed as in Proposition (1.1). First, we compute the determinant:

$$\det\ [A + \nu\cdot A_1] = \det \begin{bmatrix} -1 & \nu \\ \nu & \nu\cdot c_1^2 \end{bmatrix} = -\ \nu c_1^2 - \nu^2$$

Thus $\nu_1 = -\ c_1^2$, $\nu_2 = 0$. It follows that $y_1 = \begin{bmatrix} -c_1^2 \\ 1 \end{bmatrix}$ and $y_2 = \begin{bmatrix} 0 \\ \alpha \end{bmatrix}$ $\alpha\varepsilon R$.

Since $y_1^T\ Q_1\ y_1 = -\ c_1^2 < 0$ and $y_2^T\ Q_1\ y_2 = \alpha^2\ c_1^2 > 0$, the stabilizing controls u belong to the interval $\Gamma = (-c_1^2,\ 0)$. This can be immediately verified. Indeed, one can write:

$$x^T Qx + u \cdot x^T Q_1 x = -(x_1 - ux_2)^2 + u(c_1^2 + u)\, x_2^2 \qquad (1.21)$$

(1.21) is obviously negative for any $u \varepsilon (-c_1^2, 0)$.

Q.E.D.

The next examples are important as the following natural question arises: if a bilinear system cannot be stabilized by constant controls, under what conditions does there exist a feedback law that will stabilize it? That such a situation occurs is easy to see from the following simple numerical applications. Let:

$$\dot{x} = u \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 1 \\ 1 \end{bmatrix} v \qquad (1.22)$$

There exist no real scalars u, v that would yield x=0 as the globally asymptotically stable equilibrium point of (1.22). Yet one can check that:

$$u(x) = x_1 + x_2 - x_1x_2, \; v(x) = -x_1 - x_2 - x_1x_2 \qquad (1.23)$$

will stabilize (1.22) by using the Lyapunov function $V = \frac{1}{4} x^T \cdot x$. Likewise for:

$$x = \left(u_1 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + u_2 \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \right) x \qquad (1.24)$$

the controls $u_1(x) = -2x_1x_2$, $u_2(x) = -x_1^2 + x_2^2$ asymptotically stabilize (1.24) whereas no real constants u_1, u_2 will.

The interest in non-constant control stabilization is further motivated by the fact that the class of bilinear systems described by (1.1) is not closed under state feedback. Indeed, it is trivial to check that R.H.S. of (1.1) is linear in u_1 and x but not in both. However, if $u_1 = \phi_i(x)$ for some non-constant function ϕ_i, (1.1) will no longer be linear in x. An initial discussion concerning this problem is given in [2]. Further extensions appear in [3], [5]. Proposition 1 in [2] defines a class of state quadratic feedback controls which will stabilize the system

$$\dot{x}(t) = \left(\sum_{i=1}^{m} u_i(t) A_i \right) x + b \cdot v(t), \qquad y = C^T x \qquad (1.25)$$

where $A_i \, \varepsilon \, R^{nxn}$, b, C ε R^n, $u_i(t)$, v(t) are real scalar piecewise continuous functions. It will be interesting to find to what extent the class of controls (3) - (4) ([5], p.479) can be enlarged preserving condition (6) but eliminating (6)' ([2], p.480).

2. NON CONSTANT CONTROL STABILIZATION

Let us denote by O(f) the class of odd piecewise continuous functions on the real axis, f(•), for which $f(\sigma)=0$ only if $\sigma=0$. Note that the familiar class of "bang-bang" controls u = sgn(x) belongs to O(f). The following theorem generalizes Proposition 1 in [2].

Theorem 2.1

Consider the system (1.25). If for some negative definite matrix P, some $i_1 \varepsilon \{1, \ldots., m\}$, and for any $f_{1,2} \varepsilon O(f)$, one of the conditions below is satisfied, then (*) in Proposition 1 is verified

(a_1): $u_{i_1} = f_1(x^TQ_{i_1}x) + y$, $u_i = 0$, $i \neq i_1$, $v = x^TQ_{i_1}x - y$,

$C^T = -2b^TP$; (a_2): $u_{i_1} = f_1(x^TQ_{i_1}x)$, $u_i = 0$, $i \neq i_1$,

$v = -f_2(y)$, $C^T = -2b^TP$ or $C^T = -\frac{2}{3}b^TP$;

(a_3): $u_{i_1} = f_1(x^TQ_{i_1}x) - f_1(y)$, $v = f_1(x^TQ_{i_1}x) - q\cdot f_1(y)$,

$q > 1, f_1 \varepsilon 0(f)$, and monotone nondecreasing $C^T = -2b^TP$

Moreover, if (6) in [2] is verified then (**) is valid.

Proof. The procedure is entirely similar with the one given for the proof of Proposition 1 in [2]. Let $P = P^T$. Associating, as before, the integral: $\eta(0,t_1) = \int_0^{t_1}(v - \sum_{i=1}^{m} u_i)C^Tx$, one obtains:

$$\eta(0,t_1) = \int_0^{t_1}\{\sum_{i=1}^{m} u_i(x^TQ_ix - C^Tx) + v(2b^TP + C^T)x\}dt$$
$$-x^T(t_1)\,Px(t_1) + x^T(0)\,Px(0) \tag{2.1}$$

(a_1): Replacing the controls in (2.1), we get (for $P = P^T < 0$):

$$\beta_1||x(t_1)||^2 \leq \int_0^{t_1}\{-f_1(x^TQ_{i_1}x).x^TQ_{i_1}x - (C^Tx)^2\}\,dt + \beta_2||x(0)||^2$$

Arguing now as in (12) - (16) in [2], the result follows.

(a_2): Proceeding as above and replacing the controls, one obtains

$$\beta_1||x(t_1)||^2 \leq \int_0^{t_1}\{-f_2(C^Tx)C^Tx - f_1(x^TQ_{i_1}x).x^TQ_{i_1}x\}\,dt +$$
$$+\ \beta_2||x(0)||^2$$

The rest follows in a similar manner.

(a_3): Since $f_1(\cdot)$ is monotone nondecreasing for any $\sigma_1=\sigma_2$, $(f_1(\sigma_1)-f_1(\sigma_2))\ (\sigma_1-\sigma_2) \geq 0$. Following the same routine we get:

$$\beta_1||x(t_1)||^2 \leq \int_0^{t_1}\{-(q-1)f_1(C^Tx)C^Tx - [f_1(xTQ_{i_1}x)-f_1(C^Tx)]$$
$$[x^TQ_{i_1}x - C^Tx]\}\,dt + \beta_2||x(0)||^2$$

from which the result follows as before.

Q.E.D.

As previously mentioned, this theorem incorporates the "bang-bang" control cases. Since such controls play an important role in linear systems analysis and synthesis, it is interesting to see their impact upon stabilization in the bilinear systems set up. Moreover, due to the fact that algebraic conditions for controllability of bilinear systems of the type (1.1) with controls u_i piecewise constant and $|u_i| \leq 1$ are known (see [6]), one may wish to see to what extent the presently known controllability criteria affect feedback stabilization problems.

The next example illustrates this relation by using a simple sufficient controllability condition for (1.1), with m=2, $|u_i| \leq 1$.

C.1 (Kucera): If $B_i, \ldots, B_k$ are the basis elements for the Lie algebra $L = \{A_1, A_2\}_A$ and if rank $[B_1x, \ldots, B_kx] = n$ for all $x \varepsilon R^n - \{0\}$, then (1.1) is controllable; i.e., for any x', x" in R^n there exists a time $T \geq 0$ and a control such that if $x(0) = x'$, then $x(T) = x''$.

Example. Let:

$$x = u_1 \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + u_2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x \quad |u_1| \leq 1, \quad i = 1.2.$$

Since $[A_1, A_2] = 0$, A_1 and A_2 are the basis elements for the Lie algebra, L. Kucera's rank condition is not satisfied: rank $\begin{bmatrix} x_2 & x_1 \\ 0 & x_2 \end{bmatrix} \neq 2$ for $x_2 = 0$.

However, for $P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, one has: $x^T(A_1^T P + PA_1)x = 2\, x_1x_2$, $x^T(A_2^T P + PA_2)x = x_1^2 + x_2^2$. The controls $u_1 = -\mathrm{sgn}\, x_1x_2$, $u_2 = -1$ will stabilize the system, as can be verified by using $V(x) = x_1^2 + x_2^2$ and computing $\dot{V}(x) = - x_1x_2 \,\mathrm{sgn}\, x_1x_2 - (x_1^2 + x_2)^2 < 0$ for $x \neq 0$.

CONCLUSIONS

In this paper, we have approached the problem of stabilizing bilinear systems (I.1) by constant controls using properties associated with pairs of quadratic forms defined on real linear spaces. Those properties, described by Hestenes in [1], have been used to characterize the set U of stabilizing controls, as an open and convex set. For systems with drift term, the elements of U can be algorithmically determined. In the case of bilinear systems with non constant feedback, the class of stabilizing controls given in [2] has been enlarged to include real odd piecewise continuous functions.

REFERENCES

[1] M.R. Hestenes, "Pairs of quadratic forms", Linear Algebra and its Applications, Vol. I, No. 3, July 1968.

[2] T. Ionescu, R.V. Monopoli, "Stabilization of a class of nonlinear systems via hyperstability", 12th Annual Allerton Conference on Circuit and System Theory, October, 1974, University of Illinois, Urbana.

[3] T. Ionescu, R.V. Monopoli, "On the stabilization of bilinear systems", IEEE Transaction A-C, April 1975.

[4] W.T. Reid, "A theorem on quadratic forms", Trans. Amer. Soc. 47 (1940), pp. 437-450.

[5] T. Ionescu, "The stabilization of bilinear systems and its application to model reference adaptive systems", Ph.D. dissertation, Department of Electrical and Computer Engineering, Univ. of Mass., Amherst, 1975.

[6] D. Elliott, "Controllability criteria for bilinear systems", 12th Annual Allerton Conference on Circuit and System Theory, October, 1974, University of Illinois, Urbana.

POLE PLACEMENT DESIGN OF MULTIPLE CONTROLLER SYSTEMS VIA WEAK AND STRONG CONTROLLABILITY[†]

JOE H. CHOW
Coordinated Science Laboratory and Department of Electrical Engineering
University of Illinois at Urbana, Illinois 61801

ABSTRACT

For linear time invariant systems, the notion of weak and strong controllability is introduced using an eigenvalue sensitivity analysis. In systems with multiple controllers, the modes of the systems can be assigned to different controllers according to this relative measure of controllability. Using this decomposition, a procedure for reduced order pole placement design is proposed.

I. INTRODUCTION

The controllability conditions for a linear time-invariant system

$$\dot{x} = Ax + Bu \tag{1}$$

are well known [1,2]. In system (1), the state x is an n-vector, the input u an m-vector and A,B are n × n, n × m matrices, respectively. These conditions state whether an eigenvalue of system (1) can or cannot be affected by an arbitrary control u = Fx. However, they do not indicate how far a controllable eigenvalue will be shifted when a feedback gain matrix F of a certain magnitude is used. A preliminary attempt to relate the shift of the eigenvalues to the magnitude of F is in [3]. In this paper, an eigenvalue sensitivity [4] defines the notion of the weak and strong controllability of the eigenvalues, which can be used to investigate this relationship.

An application of the weak and strong controllability is an approximate pole placement design scheme of a multiple controller system

$$\dot{x} = Ax + \sum_{i=1}^{k} B_i u_i \tag{2}$$

where u_i, i=1,2,...,k, are the local controllers. First, the eigenvalues of system (2) are assigned to different controllers or groups of controllers according to their relative measure of controllability. This is a natural subsystem decomposition as only the properties of the system matrices A, $B_1,\ldots,B_k$ are used and it is invariant to any state transformation of system (2). Second, the pole placement of the low order subsystems designed separately is an approximate design scheme for the complete system (2). The treatment here follows the approaches of modal control [5] and decentralized control of large scale systems [6,7].

[†] This work is supported in part by the National Science Foundation under Grant ENG 74-20091, in part by Energy Research and Development Administration under Contract U.S. ERDA E(49-18)-2088, and in part by the U.S. Air Force under Grant AFOSR 73-2570.

II. EIGENVALUE SENSITIVITY

Consider the pole placement of the single input system

$$\dot{x} = Ax + bu \tag{3}$$

obtained by a special feedback control of the form

$$u = \alpha w_i' x \tag{4}$$

where w_i is the normalized adjoint eigenvector corresponding to the eigenvalue λ_i of the matrix A and α is a positive scalar. Here we assume that A possesses distinct eigenvalues.

The feedback control (4) will shift the eigenvalue λ_i to $\lambda_i + \alpha(w_i, b)$ while the rest of the eigenvalues of A remain unchanged. The symbol (w_i, b) denotes the inner product of the vectors w_i and b. In the limiting case as $\alpha \rightarrow 0^+$, the eigenvalue sensitivity of λ_i [4] is

$$S_x(\lambda_i) = \frac{\partial \lambda_i}{\partial \alpha} = (w_i, b) \ . \tag{5}$$

The subscript x in $S_x(\lambda_i)$ denotes the use of the x variable. This is important because the value of the sensitivity (5) depends on the state variables. As an illustration, for the system $\dot{x} = -x + \epsilon u$ where $0 < \epsilon \ll 1$, $|S_x(-1)| = \epsilon$. Using the transformation $y = x/\epsilon$, the system becomes $\dot{y} = -y + u$ and $|S_y(-1)| = 1$. Thus the sensitivities for different system representations can be drastically different.

Using (5), the feedback control

$$u = \left(\sum_{i=1}^{n} \frac{\Delta\lambda_i w_i'}{S_x(\lambda_i)} \right) x \tag{6}$$

for system (3) will approximately shift the eigenvalues λ_i to $\bar{\lambda}_i = \lambda_i + \Delta\lambda_i$, $i = 1,2,\ldots,n$. Here we assume that all the eigenvalues are controllable and hence their sensitivities are nonzero [2]. If $|S_x(\lambda_i)|$ is large, then only a small feedback gain vector $\Delta\lambda_i w_i'/S_x(\lambda_i)$ is necessary to shift λ_i by $\Delta\lambda_i$, and vice versa. Furthermore, the original eigenvalue locations are also important in practical situations where we are interested in the relative improvement of the system performance, which depends on the ratio $\Delta\lambda_i/\lambda_i$, the relative shift of the eigenvalues. This leads to the following definition of weak and strong controllability.

Definition: The nonzero real eigenvalue λ_i of system (3) is said to be weakly controllable if $|S_x(\lambda_i)/\lambda_i| \ll 1$ and strongly controllable if $|S_x(\lambda_i)/\lambda_i| \gtrsim 1$. The nonzero real (imaginary) part of the complex eigenvalue λ_j of system (3) is said to be weakly controllable if $|S_x(\lambda_j)/\mathrm{Re}\lambda_j| \ll 1$ ($|S_x(\lambda_j)/\mathrm{Im}\lambda_j| \ll 1$) and strongly controllable if $|S_x(\lambda_j)/\mathrm{Re}\lambda_j| \gtrsim 1$ ($|S_x(\lambda_j)/\mathrm{Im}\lambda_j| \gtrsim 1$).[†]

[†] The symbol "$\gtrsim$" means greater than or approximately equal to.

This is a logarithmic eigenvalue sensitivity definition. Thus for the case of a controllable zero eigenvalue, the sensitivity is infinite and hence it is strongly controllable. From this definition, strongly controllable eigenvalues are desirable as small feedback gains can be used to achieve big increases in system performances.

III. SYSTEMS WITH MULTIPLE CONTROLLERS

For a large scale system with multiple controllers, the eigenvalue sensitivity can be used to assign the modes of the system to different controllers or groups of controllers. Without loss of generality, let us consider a two controller system

$$\dot{x} = Ax + b_1 u_1 + b_2 u_2 \quad . \tag{7}$$

The results obtained for system (7) can be easily extended to systems with more than two controllers. Here we assume that the modes of A are distinct and can be controlled by u_1 or u_2 or both of them. In [6], system (7) is decomposed according to the controllable subspaces of u_1 and u_2. In the following, system (7) will be decomposed according to the weakly and strongly controllable subspaces of u_1 and u_2.

For system (7), the eigenvalue sensitivities of λ_i with respect to u_1 and u_2 are $S_{x1}(\lambda_i) = (w_i, b_1)$ and $S_{x2}(\lambda_i) = (w_i, b_2)$, respectively. The mode λ_i is assigned to u_1 or u_2 depending on its relative weak and strong controllability with respect to the controllers. If $|S_{x1}(\lambda_i)| \gg |S_{x2}(\lambda_i)|$, the mode λ_i is more sensitive to a feedback control through u_1 than through u_2 and hence we say that the mode λ_i is under the control of u_1. Similarly, if $|S_{x2}(\lambda_i)| \gg |S_{x1}(\lambda_i)|$, then λ_i is under the control of u_2. If $|S_{x1}(\lambda_i)| \simeq |S_{x2}(\lambda_i)|$, then λ_i is under the control of both u_1 and u_2.

From our previous discussion, we know that the values of $|S_{x1}(\lambda_i)|$ and $|S_{x2}(\lambda_i)|$ will be different under a different representation of the system (7). However, the relative magnitude of the sensitivity with respect to u_1 and u_2 will remain the same for any nonsingular transformation.

Theorem 1

The relative ratio of the eigenvalue sensitivity of the system (7) with respect to u_1 and u_2 is invariant to any nonsingular transformation of system (7).

Proof: A transformation $x = Ty$ of system (7) yields

$$\dot{y} = T^{-1}ATy + T^{-1}b_1 u_1 + T^{-1}b_2 u_2 \quad . \tag{8}$$

Then the sensitivities of λ_i with respect to u_1 and u_2 are $S_{y1}(\lambda_i) = (\overline{w}_i, T^{-1}b_1)$ and $S_{y2}(\lambda_i) = (\overline{w}_i, T^{-1}b_2)$, respectively, where $\overline{w}_i = \pm T'w_i / |T'w_i|$. Thus

$$\frac{S_{y1}(\lambda_i)}{S_{y2}(\lambda_i)} = \frac{(T'w_i, T^{-1}b_1)/|M'w_i|}{(T'w_i, T^{-1}b_2)/|M'w_i|} = \frac{S_{x1}(\lambda_i)}{S_{x2}(\lambda_i)} \tag{9}$$

which proves the theorem.

Theorem 1 implies that if for a certain mode λ_i, $|S_{x1}(\lambda_i)| \gg |S_{x2}(\lambda_i)|$, then under all the representations of the system, λ_i is always more sensitive to u_1 than u_2. Thus, a certain mode will belong to the same controller no matter which set of state variables is being used. Since some eigenvalues are more sensitive to one controller than the other, a subsystem decomposition scheme can be used for the pole placement design of system (7).

IV. NONINTERACTIVE CONTROLLERS

First we assume that system (7) possesses the following property

$$\text{(I)} \quad |S_{x1}(\lambda_i)| \gg |S_{x2}(\lambda_i)|, \; |S_{x1}(\lambda_j)| \ll |S_{x2}(\lambda_j)|, \; i=1,2,\ldots,k, \; j=k+1, k+2,\ldots,n,$$

that is, the modes λ_i and λ_j can be distinctly assigned to controllers u_1 and u_2, respectively. Here there are predominantly two subsystems and for practical purposes, the controllers are noninteractive.

A model transformation $y_1 = W_1x$, $y_2 = W_2x$, where the rows of W_1, W_2 are the adjoint eigenvectors corresponding to λ_i, λ_j, respectively, yields the modal system of (7)

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} W_1b_1 \\ W_2b_1 \end{bmatrix} u_1 + \begin{bmatrix} W_1b_2 \\ W_2b_2 \end{bmatrix} u_2 \tag{10}$$

where $\Lambda_1 = \text{diag}\,(\lambda_1,\ldots,\lambda_k)$, $\Lambda_2 = \text{diag}\,(\lambda_{k+1},\ldots,\lambda_n)$. Theorem 1 and property (I) guarantee that each element in W_1b_1 and W_2b_2 is much larger than the corresponding elements in W_1b_2 and W_2b_1 in magnitude, and system (10) can be approximated by the subsystems

$$\dot{\bar{y}}_1 = \Lambda_1\bar{y}_1 + W_1b_1\,u_1 \tag{11}$$

$$\dot{\bar{y}}_2 = \Lambda_2\bar{y}_2 + W_2b_2\,u_2 \quad . \tag{12}$$

The pole placement design of these subsystems can be accomplished by the controls $u_1 = f_1\bar{y}_1$ and $u_2 = f_2\bar{y}_2$. This is an approximate pole placement design for the complete system as shown in the following theorem.

Theorem 2

The controls $u_1 = f_1W_1x$, $u_2 = f_2W_2x$ can place the poles of system (7) possessing property (I) arbitrarily close to the poles of the matrices $\bar{\Lambda}_1 = \Lambda_1 + W_1b_1f_1$ and $\bar{\Lambda}_2 = \Lambda_2 + W_2b_2f_2$.

Proof: Using the transformation $y_1 = W_1x$, $y_2 = W_2x$, the system matrix of system (7) controlled by $u_1 = f_1W_1x$, $u_2 = f_2W_2x$ becomes

$$\begin{bmatrix} \bar{\Lambda}_1 & W_1 b_2 f_2 \\ W_2 b_1 f_1 & \bar{\Lambda}_2 \end{bmatrix} . \tag{13}$$

Since the elements of the matrices $W_2 b_1 f_1$ and $W_1 b_2 f_2$ are much smaller than those of $\bar{\Lambda}_1$ and $\bar{\Lambda}_2$ in magnitude, respectively, the eigenvalues of matrix (13) are arbitrarily close to the eigenvalues of $\bar{\Lambda}_1$ and $\bar{\Lambda}_2$.

This approximate pole placement scheme has been proposed for systems with weak coupling

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} A_1 & \varepsilon A_2 \\ \varepsilon A_3 & A_4 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} b_1 \\ \varepsilon b_2 \end{bmatrix} u_1 + \begin{bmatrix} \varepsilon b_3 \\ b_4 \end{bmatrix} u_2 \tag{14}$$

where ε is a small positive scalar. Weak coupling in system (7) is in general not easy to recognize, but property (I) guarantees that it can be transformed into (10) which is a special form of system (14). Thus we have proposed a more general criterion for a system to be weakly coupled.

V. TWO STAGE INTERACTIVE DESIGN

A more general situation occurs when there is an interaction between the controllers u_1 and u_2 in system (7), that is, it possesses the property

(II) $$|s_{x1}(\lambda_i)| \gg |s_{x2}(\lambda_i)| , \quad i=1,2,\ldots,n_1$$

$$|s_{x1}(\lambda_j)| \ll |s_{x2}(\lambda_j)| , \quad j=n_1+1,\ldots,n_2$$

$$|s_{x1}(\lambda_k)| \simeq |s_{x2}(\lambda_k)| , \quad j=n_2+1,\ldots,n \ .$$

A two stage pole placement design for system (7) possessing property (II) is proposed as follows. For notational simplicity, let W_1, W_2, W_3 be matrices whose rows are the adjoint eigenvectors corresponding to λ_i, λ_j, λ_k, respectively.

The first stage of the design consists of designing $u_1 = f_1 W_1 x + v_1$, $u_2 = f_2 W_2 x + v_2$ such that the feedback system

$$\dot{x} = (A + b_1 f_1 W_1 + b_2 f_2 W_2)\, x + b_1 v_1 + b_2 v_2 \tag{15}$$

possesses n_2 eigenvalues arbitrarily close to those of $\bar{\Lambda}_1 = \Lambda_1 + b_1 f_1$ and $\bar{\Lambda}_2 = \Lambda_2 + b_2 f_2$, where $\Lambda_1 = \text{diag}\,(\lambda_1,\ldots,\lambda_{n_1})$ and $\Lambda_2 = \text{diag}\,(\lambda_{n_1+1},\ldots,\lambda_{n_2})$.

The first stage design changes the adjoint eigenvectors of λ_k, which now satisfy the equation

$$(W_3 + C_1 W_1 + C_2 W_2)(A + b_1 f_1 W_1 + b_2 f_2 W_2) = \Lambda_3 (W_3 + C_1 W_1 + C_2 W_2) \tag{16}$$

where $\Lambda_3 = \mathrm{diag}\,(\lambda_{n_2+1},\ldots,\lambda_n)$ and the rows of the matrix $W_3 + C_1W_1 + C_2W_2$ are the new adjoint eigenvectors. Using property (II), (16) can be approximated by

$$C_{10}\bar{\Lambda}_1W_1 + C_{20}\bar{\Lambda}_2W_2 + W_3b_1f_1W_1 + W_3b_2f_2W_2 = \Lambda_3C_{10}W_1 + \Lambda_3C_{20}W_2 \tag{17}$$

where C_{10}, C_{20} are approximations of C_1, C_2. Equation (17) will be satisfied if C_{10} and C_{20} satisfy the equations

$$C_{10}\bar{\Lambda}_1 - \Lambda_3C_{10} = -W_3b_1f_1\ , \quad C_{20}\bar{\Lambda}_2 - \Lambda_3C_{20} = -W_3b_2f_2\ . \tag{18}$$

Note that C_1 and C_2 are coupled in (16), but C_{10} and C_{20} are decoupled in (18). C_{10} and C_{20} can be solved from (18) provided that the eigenvalues of $\bar{\Lambda}_1$ and $\bar{\Lambda}_2$ are different from those of Λ_3 [8].

Now the transformation

$$y_1 = W_1x,\quad y_2 = W_2x,\quad y_3 = W_{30}x \tag{19}$$

where $W_{30} = W_3 + C_{10}W_1 + C_{20}W_2$, for system (15) gives

$$\begin{bmatrix}\dot{y}_1\\ \dot{y}_2\\ \dot{y}_3\end{bmatrix} = \begin{bmatrix}\bar{\Lambda}_1 & W_1b_2f_2 & 0\\ W_2b_1f_1 & \bar{\Lambda}_2 & 0\\ C_{20}W_2b_1f_1 & C_{10}W_1b_2f_2 & \Lambda_3\end{bmatrix}\begin{bmatrix}y_1\\ y_2\\ y_3\end{bmatrix} + \begin{bmatrix}W_1b_1\\ W_2b_1\\ W_{30}b_1\end{bmatrix}v_1 + \begin{bmatrix}W_1b_2\\ W_2b_2\\ W_{30}b_2\end{bmatrix}v_2. \tag{20}$$

The dynamics of y_3 can be approximated by

$$\dot{\bar{y}}_3 = \Lambda_3\bar{y}_3 + W_{30}b_1v_1 + W_{30}b_2v_2 \tag{21}$$

and hence the second stage of the design is to implement $v_1 = f_3\bar{y}_3$, $v_2 = f_4\bar{y}_3$ such that the matrix $\bar{\Lambda}_3 = \Lambda_3 + W_{30}b_1f_1 + W_{30}b_2f_2$ has the desired equivalues.

Theorem 3

If the controls $u_1 = (f_1W_1 + f_3W_{30})x$, $u_2 = (f_2W_2 + f_4W_{30})x$ are implemented to the system (7) possessing property (II), then the eigenvalues of the feedback system are arbitrarily close to those of the matrices $\bar{\Lambda}_1$, $\bar{\Lambda}_2$, $\bar{\Lambda}_3$.

The proof of Theorem 3 is similar to that of Theorem 2 except that the transformation (19) is used.

The advantage of this two stage design over the modal design in [5] is that separate subsystems are designed in the first stage, and a simplified procedure is used to find the new adjoint eigenvectors in the second stage. Computation effort is reduced in both stages.

VI. EXAMPLE

A dynamic model of a VSTOL aircraft is given by [5]

$$\dot{x} = \begin{bmatrix} -.0506 & 0 & -1 & .2380 \\ -.7374 & -1.3345 & .3696 & 0 \\ .0100 & .1074 & -.3320 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} x + \begin{bmatrix} .0409 \\ 1.2714 \\ -2.0625 \\ 0 \end{bmatrix} u_1 + \begin{bmatrix} 0 \\ -20.3106 \\ 1.3350 \\ 0 \end{bmatrix} u_2 \tag{22}$$

where the states x_1 is the sideslip angle, x_2 the roll rate, x_3 the yaw rate, x_4 the bank angle and the controls u_1 and u_2 are the rudder angle and the aileron angle, respectively. The eigenvalues of the system are $\lambda_1 = -.4023$, $\lambda_2 = -1.4088$, $\lambda_{3,4} = .0470 \pm j.3147$, corresponding to the rolling, the spiral and the unstable oscillatory Dutch-roll mode of the aircraft, respectively. The sensitivities are found to be $S_{u1}(\lambda_1) = -1.9718$, $S_{u2}(\lambda_1) = .1724$, $S_{u1}(\lambda_2) = -.8366$, $S_{u2}(\lambda_2) = 17.444$, $S_{u1}(\lambda_{3,4}) = -1.8147 \pm j.2081$ and $S_{u2}(\lambda_{3,4}) = 1.7856 \mp j3.4595$. Hence λ_1 should be controlled by u_1, λ_2 by u_2 and $\lambda_{3,4}$ by both u_1 and u_2.

In the first stage design, f_1 and f_2 are computed to shift λ_1 to -.8 and λ_2 to -2.8, and if the feedback

$$u = \begin{bmatrix} f_1 W_1 \\ f_2 W_2 \end{bmatrix} x = \begin{bmatrix} .0183 & .0115 & .2003 & -.0108 \\ .0375 & .0692 & .0111 & -.0063 \end{bmatrix} x \tag{23}$$

is implemented to system (22), the eigenvalues are found to be -.7989, -2.8013, .0470 ± j.3147.

In the second stage, f_4 is computed to shift $\lambda_{3,4}$ to $-.4 \pm j1$ (u_3 not being used), and if the feedback

$$u = \begin{bmatrix} f_1 W_1 \\ f_2 W_2 + f_4 W_{30} \end{bmatrix} x = \begin{bmatrix} .0183 & .0115 & .2003 & -.0108 \\ .6799 & .0579 & -.8306 & .1637 \end{bmatrix} x \tag{24}$$

is implemented to system (22), the eigenvalues are found to be -.7992, -2.755, -.4230 ± j.9259. In both designs, the actual eigenvalues are close to the desired ones.

VII. CONCLUSION

Weak and strong controllability is introduced through an eigenvalue sensitivity analysis. The relative weak and strong controllability is extended to systems with two controllers to assign the modes of the systems to the controllers. This assignment is invariant for all state transformations of the system. A subsystem pole placement scheme is proposed for noninteractive controllers, and a two stage design scheme is proposed for interactive controllers. In both cases, the feedback controls so obtained place the eigenvalues of the complete system arbitrarily close to those desired.

ACKNOWLEDGEMENT

The author is indebted to Professor P. V. Kokotovic for his helpful comments.

REFERENCES

[1] C.-T. Chen and C. A. Desoer, "Controllability and Observability of Composite Systems," IEEE Trans. on Automatic Control, Vol. AC-12, No. 4, August 1967, pp. 402-409.

[2] M. L. J. Hautus, "Controllability and Observability Conditions of Linear Autonomous Systems," Proc. Kon. Neb. Akad. Wetensch., Ser. A, Vol. 72, No. 5, 1969, pp. 443-448.

[3] J. H. Chow, "Preservation of Controllability in Linear Time-Invariant Perturbed Systems," to appear in Int. J. Control.

[4] J. E. Van Ness, J. M. Boyle and F. P. Imad, "Sensitivities of Large, Multiple-Loop Systems," IEEE Trans. on Automatic Control, Vol. AC-10, No. 3, July 1965, pp. 308-315.

[5] B. Porter and R. Crossley, Modal Control: Theory and Applications, Taylor and Francis, London, 1972.

[6] M. Aoki, "On Feedback Stabilization of Decentralized Dynamic Systems," Automatica, Vol. 8, 1972, pp. 163-173.

[7] S.-H. Wang and E. J. Davison, "On the Stabilization of Decentralized Control Systems," IEEE Trans. on Automatic Control, Vol. AC-18, No. 5, October 1973, pp. 473-478.

[8] M. C. Pease III, Methods of Matrix Algebra, Academic Press, New York 1965.

A CONDITION FOR DECENTRALIZATION IN MODEL REFERENCE ADAPTIVE SYSTEMS

MALUR K. SUNDARESHAN
Department of Electrical Engineering
Concordia University
Montreal, Canada.

ABSTRACT

It is shown that when the interconnection pattern within a system conforms to certain prescribed symmetry requirements, decentralized adaptive algorithms for identification and control can be developed within a decomposition-decentralization framework. This is useful in the presence of constraints on the available information structure of the system.

I. INTRODUCTION

Over the past decade, methods based on Lyapunov synthesis have emerged as quite powerful tools for the design of model reference adaptive systems and many new algorithms for both identification and control [1-4] have been developed. A central feature of these schemes is that their implementation requires the entire state vector of the plant to be available for measurement and monitoring. Hence a straightforward application of these will not, in general, be feasible in the case of large-scale systems, either due to the presence of constraints on the available information structure [5] which do not permit a centralized information of the total state of the system or due to the fact that the implementation of a single adaptive scheme may require the simultaneous adjustment of a large number of parameters to assure the stability of the overall adaptive system. Moreover, many of the large systems occurring in practice can be identified as interconnections of several subunits and it is useful to take advantage of the decomposed system structure available. In this note, we shall present a simple condition which ensures the synthesis of adaptive systems to be conducted within a decomposition-decentralization framework. The principal philosophy behind the method of analysis adopted here is the characterization of the interconnections within the system as perturbation signals acting in contradiction to the autonomy of the individual subsystems. For technical reasons, the details of discussion are given with respect to the adaptive identification problem; the corresponding development for the control problem, as is well-known, runs parallel.

II. PROBLEM DESCRIPTION

Consider a large-scale system $\mathcal{S}$ that may be described as an interconnection of s subsystems $\mathcal{S}_1, \mathcal{S}_2, \dots \mathcal{S}_s$ by

$$\dot{x}_i(t) = A_i x_i(t) + B_i u_i(t) + \sum_{j=1}^{s} H_{ij}(t)\, x_j(t), \quad i=1, 2, \dots s, \qquad (1)$$

where $x_i(.) : \mathcal{R} \to \mathcal{R}^{n_i}$ is the state of the i-th subsystem $\mathcal{S}_i$ and $u_i(.) : \mathcal{R} \to \mathcal{R}^{m_i}$, its control function. A_i and B_i are constant matrices of appropriate dimensions and $H_{ij}(.) : \mathcal{R} \to \mathcal{R}^{n_i x n_j}$ specifies the interaction from $\mathcal{S}_j$ to $\mathcal{S}_i$. It may be noted that $H_{ij}(.)$ are elements of the nxn matrix $H(.) = [H_{ij}(.)]$, i,j=1, 2, ... s, $n = \sum_{i=1}^{s} n_i$, that describes the interconnection pattern within the overall system $\mathcal{S}$.

We shall assume that the elements of A_i and B_i are unknown while the elements of H(.) are known exactly[†]. We shall also assume the presence of a constraint on the available information structure of the system, that the state $x_i(.)$ of subsystem $\mathcal{S}_i$ is available for measurement only at $\mathcal{S}_i$ and a centralized information of the total state is not permitted. Furthermore, while each of the subsystems is known to be completely controllable in isolation (i.e. under conditions of decoupling, $H_{ij}(.) \equiv 0$), the overall system $\mathcal{S}$ may or may not be completely controllable.

The problem of interest now is to determine conditions under which simple adaptive algorithms for the identification of the subsystem matrices A_i and B_i can be developed.

III. DEVELOPMENT OF THE DECENTRALIZED ADAPTIVE SCHEME

Subsystem $\mathcal{S}_i$, when assumed to be decoupled from the rest of the system, is described by

$$\dot{x}_i(t) = A_i x_i(t) + B_i u_i(t) \tag{2}$$

and hence, following [2], adaptive laws for the identification of A_i and B_i with the available information structure can be developed by choosing a model

$$\dot{y}_i(t) = C_i y_i(t) + [\tilde{A}_i(t) - C_i]\, x_i(t) + \tilde{B}_i(t)\, u_i(t) \tag{3}$$

where $C_i \varepsilon \mathcal{R}^{n_i x n_i}$ has all eigenvalues in the open left-half plane. The adaptive updating equations are given by,

$$\begin{aligned} &\dot{\tilde{A}}_i(t) = -\Gamma_i P_i e_i(t)\, x_i^T(t) \\ \text{and}\quad &\dot{\tilde{B}}_i(t) = -\Lambda_i P_i e_i(t)\, u_i^T(t) \end{aligned} \tag{4}$$

[†] Such an assumption is not unrealistic since in many of the practical large-scale systems, the interconnection elements which are gains determining the magnitude of information flow among the various subunits will be known, whereas an exact knowledge of the dynamics of these processes themselves will not be available.

where $e_i(.) : \mathcal{R} \to \mathcal{R}^{n_i}$ is the error vector $\{y_i(.) - x_i(.)\}$, $\Gamma_i \varepsilon \mathcal{R}^{n_i x n_i}$ and $\Lambda_i \varepsilon \mathcal{R}^{n_i x n_i}$ are arbitrary symmetric and positive definite matrices and $P_i \varepsilon \mathcal{R}^{n_i x n_i}$ is the symmetric and positive definite solution of the Lyapunov matrix equation

$$C_i^T P_i + P_i C_i + Q_i = 0 \tag{5}$$

for an arbitrary symmetric and positive definite matrix $Q_i \varepsilon \mathcal{R}^{n_i x n_i}$. As shown in [2], when (a) $u_i(t)$ is periodic, (b) the elements of $u_i(t)$ are linearly independent and (c) each element contains at least $(n_i+1)/2$ different frequency components, the model (3) asymptotically tracks the decoupled subsystems (2), i.e. $\lim_{t\to\infty} \tilde{A}_i(t) = A_i$, $\lim_{t\to\infty} \tilde{B}_i(t) = B_i$ and $\lim_{t\to\infty} e_i(t)=0$.

Assuming that the above analysis is conducted for each of the decoupled subsystems, conditions under which identification of the subsystems can be achieved when they are part of the overall system $\mathcal{S}$, are given in the following theorem.

THEOREM: Let the interconnection matrix H(t) be factorizable as

$$H(t) = [U(t) - S(t)]\, P \tag{6}$$

where $U(.) : \mathcal{R} \to \mathcal{R}^{nxn}$ is an arbitrary skew-symmetric matrix, $S(.) : \mathcal{R} \to \mathcal{R}^{nxn}$ is an arbitrary symmetric and nonnegative definite matrix and $P = \text{diag}\,[P_1\ P_2 \ldots P_s]$. Then the model

$$\dot{y}_i(t) = C_i y_i(t) + [\tilde{A}_i(t) - C_i]\, x_i(t) + \tilde{B}_i(t) u_i(t) + \sum_{j=1}^{s} H_{ij}(t) y_j(t), \quad i=1, 2, \ldots s, \tag{7}$$

with the adaptive laws (4), asymptotically tracks the overall system $\mathcal{S}$, i.e. $\lim_{t\to\infty} \tilde{A}_i(t) = A_i$, $\lim_{t\to\infty} \tilde{B}_i(t) = B_i$ and $\lim_{t\to\infty} [y_i(t) - x_i(t)] = 0 \ \forall\ i=1, 2, \ldots s$, when the inputs $u_i(t)$ are chosen as described earlier.

Proof: From (1) and (7), the error system may be obtained as

$$\dot{e}_i(t) = C_i e_i(t) + \Phi_i(t) x_i(t) + \Psi_i(t)\, u_i(t) + \sum_{j=1}^{s} H_{ij}(t)\, e_j(t), \quad i=1, 2, \ldots s, \tag{8}$$

where $\Phi_i(t) = \{\tilde{A}_i(t)-A_i\}$ and $\Psi_i(t) = \{\tilde{B}_i(t)-B_i\}$ are the parameter errors. (8) may be equivalently described by the composite equation

$$\dot{e}(t) = (C+H)\ e(t) + \Phi(t)\ x(t) + \Psi(t)\ u(t) \tag{9}$$

where $e(.) : \mathcal{R} \to \mathcal{R}^n$, $x(.) : \mathcal{R} \to \mathcal{R}^n$ and $u(.) : \mathcal{R} \to \mathcal{R}^m$ are the composite vectors $e^T = [e_1^T\ e_2^T\ \dots\ e_s^T]$, $x^T = [x_1^T\ x_2^T\ \dots\ x_s^T]$, $u^T = [u_1^T\ u_2^T\ \dots\ u_s^T]$ and $C = \text{diag}\ [C_1\ C_2\ \dots\ C_s]$, $\Phi = \text{diag}\ [\Phi_1\ \Phi_2\ \dots\ \Phi_s]$, $\Psi = \text{diag}\ [\Psi_1\ \Psi_2\ \dots\ \Psi_s]$.

Now, choose

$$V(e) = e^T Pe + \text{tr}\ [\Phi^T\Gamma^{-1}\Phi + \Psi^T\Lambda^{-1}\Psi], \tag{10}$$

where $P = \text{diag}\ [P_1\ P_2\ \dots\ P_s]$, $\Gamma = \text{diag}\ [\Gamma_1\ \Gamma_2\ \dots\ \Gamma_s]$ and $\Lambda = \text{diag}\ [\Lambda_1\ \Lambda_2\ \dots\ \Lambda_s]$, as a Lyapunov function candidate. Since each of the matrices P_i, Γ_i and Λ_i is positive definite, $V(e)$ is positive definite. Also,

$$\begin{aligned}\dot{V}(e) &= e^T\ [(C+H)^T P+P\ (C+H)] + 2\ \text{tr}\ [\Phi^T(Pex^T + \Gamma^{-1}\dot{\Phi}) \\ &\qquad + \Psi^T(Peu^T + \Lambda^{-1}\dot{\Psi})] \\ &= -\ e^T\ [Q + 2PSP]\ e \end{aligned} \tag{11}$$

where $Q = \text{diag}\ [Q_1\ Q_2\ \dots\ Q_s]$, using (4), (5) and (6) in the simplification. Since each Q_i is positive definite and S is nonnegative definite, $\dot{V}(e)$ is negative definite on the noncompact manifold of e. Hence stability of the error system (8) is ensured.

However, with the prescribed conditions on the inputs, it can be shown, following the arguments in [2], that the error system is indeed uniformly asymptotically stable in the (e, Φ, Ψ) space and this completes the proof of the theorem.

IV. CONCLUDING REMARKS

1. The result presented in the previous section identifies a condition on the interconnection pattern within the system, under which a decentralized adaptive identification scheme can be developed for the large-scale system under consideration. It may be noted that all the necessary

computations are conducted with respect to the subsystems which are of small dimensions, thus leading to minimal data-processing requirements.

2. The close relation between the condition (6) on the interconnections and the constraints on the information structure of the system is worth emphasizing. It is simple to observe that if centralized information of the total state of the system is available, then one can use a model

$$\dot{y}_i(t) = C_i y_i(t) + [\tilde{A}_i(t)-C_i]\, x_i(t) + \tilde{B}_i u_i(t) + \sum_{j=1}^{s} H_{ij}(t) x_j(t), \qquad i=1,\ 2,\ \dots\ s,$$

in lieu of (7), which asymptotically tracks the system $\mathcal{S}$ for any arbitrary interconnection pattern.

3. The present result presupposes the availability of the state $x_i(.)$ at the subsystem $\mathcal{S}_i$. However, if this is not the case, then s independent local adaptive observers may be designed for the subsystems for implementing the identification laws [2, 4, 6]. This possibility of conducting state observation within the same decomposition-decentralization framework offers a distinct advantage when the use of a single observer is not feasible owing to the largeness of the system.

4. It is well-known that the degree of negative definiteness of $\dot{V}$ determines the speed of adaptive response in Lyapunov-designed adaptive systems. The presence of an additional term involving the matrix S in (11) hence suggests that the convergence of the overall adaptive scheme is at least as fast as that of the adaptive laws developed for the decoupled subsystems. In this context, it is also worth mentioning that the Lyapunov function (10) used here is only a special case of the generalized functions developed recently by Carroll [7], which have been shown to result in faster adaptive algorithms; extension of the present work making use of these Lyapunov functions is, however, trivial.

REFERENCES

1. R.V. Monopoli, "Adaptive control and identification via Lyapunov's direct method", Proc. 1972 Annual Allerton Conf., Univ. Ill., Urbana, 1972.

2. K.S. Narendra and P. Kudva, "Stable adaptive schemes for system identification and control - Parts I and II", IEEE Trans. Systems, Man and Cybernetics, Vol. SMC-4, pp. 542-560, 1974.

3. R.V. Monopoli, "Model reference adaptive control with an augmented error signal", IEEE Trans. Auto. Control, Vol. AC-19, pp. 474-484, 1974.

4. G. Luders and K.S. Narendra, "Stable adaptive schemes for state estimation and identification of linear systems", IEEE Trans. Auto. Control, Vol. AC-19, pp. 841-847, 1974.

5. Y.C. Ho and K.C. Chu, "Information structure in dynamic multi-person control problems", Automatica, Vol. 10, pp. 341-351, 1974.

6. R.L. Carroll and D.P. Lindorff, "An adaptive observer for single-input single-output linear systems", IEEE Trans. Auto. Control, Vol. AC-18, pp. 428-435, 1973.

7. R.L. Carroll, "New adaptive algorithms in Lyapunov synthesis", IEEE Trans. Auto. Control, Vol. AC-21, pp. 246-249, 1976.

INPUT-OUTPUT STABILITY OF INTERCONNECTED SYSTEMS DESCRIBED BY NONLINEAR STOCHASTIC DIFFERENTIAL EQUATIONS*

R. L. GUTMANN
Electrical Engineering Department
and Engineering Research Institute
Iowa State University
Ames, Iowa 50011

A. N. MICHEL
Electrical Engineering Department
and Engineering Research Institute
Iowa State University
Ames, Iowa 50011

ABSTRACT

Stochastic input-output stability is considered for composite systems described by nonlinear differential equations with stochastic driving functions, random coefficients, and random initial conditions. The techniques presented apply to large scale or composite systems in the sense that stability is established by analysis of lower order subsystems and the interconnecting structure. Graphical techniques of the circle type are emphasized.

I. INTRODUCTION

In this paper we investigate the stochastic input-output stability (almost sure L_2 stability and stochastic absolute stability) of interconnected systems described by nonlinear differential equations with stochastic driving functions, random coefficients, and random initial conditions. We determine the stability of high dimensional systems from properties of lower order (and simpler) subsystems and the system interconnecting structure. Similar techniques have recently been applied to deterministic systems by Porter and Michel [1], Lasley and Michel [2,3], and Miller and Michel [4,5].

The main results of this paper are based on a theorem established in [6] which applies to interconnected systems described by nonlinear stochastic Volterra integral equations. This result has graphical interpretations of the circle type.

The stochastic nonlinear differential equations studied herein are similar to those investigated by Tsokos [7] and Tsokos and Padgett [8]. Such systems of equations can be used in the modeling of large networks, economic systems, political systems, ecological systems, biological systems, societal systems, and the like.

II. PRELIMINARIES

Let $A=[a_{ij}]$ be an n×m matrix with transpose A^T and complex conjugate transpose A^*. The inverse of a nonsingular n×n matrix A is denoted by A^{-1}. If C and D are real n×m matrices, then $C\geq D$ signifies that $c_{ij}\geq d_{ij}$ for all i and j and $C\geq 0$ means $c_{ij}\geq 0$ for all i and j. Let $I_{(N)}$ by the N×N identity matrix and let $\Lambda(M)$ denote the positive square root of the largest eigenvalue of M^*M. If the elements of a real matrix B depend on a real parameter t we say that B is bounded if there exists a real number K such that $|b_{ij}(t)|\leq K<\infty$ for all i and j and for all allowable t. Let $R=(-\infty,\infty)$, $R^+=[0,\infty)$, and $R^N=R\times R\times\cdots\times R$. If $x=[x^{(1)},x^{(2)},\cdots,x^{(N)}]^T\epsilon R^N$, then $|x|=[(x^{(1)})^2+\cdots+(x^{(N)})^2]^{\frac{1}{2}}$. The set of all real Lebesgue measurable N-vector

---

[*] This work was supported by the National Science Foundation under Grant ENG 75-14093 and by the Engineering Research Institute, Iowa State University.

valued functions of the variable $t\varepsilon R^+$ is denoted by $H_{(N)}(R^+)$. Also, $L_{p(N)}(R^+)=\{f\varepsilon H_{(N)}(R^+):\int_0^\infty|f(t)|^p dt<\infty\}$, $1\leq p<\infty$. If $f,g\varepsilon L_{2(N)}(R^+)$, then

$$\langle f,g\rangle = \int_0^\infty f^T(t)g(t)dt \text{ and } \|f\| = \langle f,f\rangle^{\frac{1}{2}}.$$

Furthermore, the norm of a linear operator T on $L_{2(N)}(R^+)$ is denoted by $\|T\|$. If $A(t)=[a_{ij}(t)]$ is an arbitrary real $N_1\times N_2$ matrix-valued Lebesgue-measurable function of $t\varepsilon R^+$, we say that $A\varepsilon K_{p(N_1\times N_2)}(R^+)$, $1\leq p<\infty$, if $\int_0^\infty|a_{ij}(t)|^p dt<\infty$ for all i and j.

Given a probability space (Ω,F,P), we denote by $X_{(N)}$ the space of all N-dimensional real-valued random vectors over Ω which are square integrable with respect to P, i.e., if $x(\omega)=[x^{(1)}(\omega),\cdots,x^{(N)}(\omega)]^T\varepsilon X_{(N)}$, then $x^{(i)}(\omega)$ is F-measurable, $i=1,\cdots,N$, and $\int_\Omega x^T(\omega)x(\omega)dP(\omega)<\infty$. Let $H_{(N)}(R^+,\Omega)$ denote the space of all real-valued N-dimensional random vector processes over $R^+\times\Omega$ such that if $x\varepsilon H_{(N)}(R^+,\Omega)$, then $x(\cdot,\omega)\varepsilon H_{(N)}(R^+)$ for fixed $\omega\varepsilon\Omega$, and $x(t,\cdot)\varepsilon X_{(N)}$ for fixed $t\varepsilon R^+$. Let $B_{(N)}(\Omega)$ denote the set

$$B_{(N)}(\Omega) = \{x\varepsilon X_{(N)}:\underset{\omega\varepsilon\Omega}{\text{ess sup}}|x(\omega)|<\infty\}$$

and define $L_{p(N)}(R^+, L_\infty(\Omega))$ by

$$L_{p(N)}(R^+, L_\infty(\Omega)) = \{x\varepsilon H_{(N)}(R^+,\Omega):\underset{\omega\varepsilon\Omega}{\text{ess sup}}\int_0^\infty|x(t,\omega)|^p dt<\infty\},$$

$1\leq p<\infty$. Let $A(t,\omega)=[a_{ij}(t,\omega)]$ denote an arbitrary real $N_1\times N_2$-matrix valued stochastic process with $a_{ij}\varepsilon H_{(1)}(R^+,\Omega)$. Then $K_{p(N_1\times N_2)}(R^+, L_\infty(\Omega))$, $1\leq p<\infty$, denotes matrix processes $A(t,\omega)$ such that

$$\underset{\omega\varepsilon\Omega}{\text{ess sup}}\int_0^\infty|a_{ij}(t,\omega)|^p dt<\infty \quad \text{for all } i \text{ and } j.$$

Let $T\varepsilon R^+$ and define for $x\varepsilon H_{(N)}(R^+,\Omega)$,

$$x_T(t,\omega)=\begin{cases} x(t,\omega), & 0\leq t\leq T \\ 0 & , t>T \end{cases}.$$

Define $E_{p(N)}=\{x\varepsilon H_{(N)}(R^+,\Omega):x_T\varepsilon L_{p(N)}(R^+, L_\infty(\Omega))\}$, $1\leq p<\infty$. Also, define the truncation operator π_T on $H_{(N)}(R^+,\Omega)$ by $\pi_T x=x_T$. Let $E_{s(N)}$ denote those processes in $E_{2(N)}$ having time derivatives a.e. $[\mu]$, where μ denotes the Lebesgue measure.

Let $\eta_{(N)}$ denote the collection of memoryless nonlinearities of the type

$$\psi(x(t,\omega), t, \omega) = [\psi^{(1)}(x^{(1)}(t,\omega), t, \omega),\cdots,\psi^{(N)}(x^{(N)}(t,\omega), t, \omega)]^T,$$

with $x=[x^{(1)},\cdots,x^{(N)}]^T\varepsilon H_{(N)}(R^+,\Omega)$, $t\varepsilon R^+$, $\omega\varepsilon\Omega$, where $\psi^{(i)}(g,t,\omega)$, $i=1,\cdots,N$, are real-valued functions of the real variables $g\varepsilon R$ and $t\varepsilon R^+$ and the variable $\omega\varepsilon\Omega$ such that

(i) $P\{\omega:\psi^{(i)}(0,t,\omega)=0,\ t\varepsilon R^+,\ i=1,\cdots, N\}=1$;

(ii) there exist real numbers a and b such that

$$P\{\omega:a\leq\frac{\psi^{(i)}(g,t,\omega)}{g}\leq b,\ t\varepsilon R^+,\ g\neq 0,\ i=1,\cdots, N\}=1;$$

(iii) $\psi^{(i)}(x(t,\omega), t,\omega)$ is a Lebesgue-measurable function of t and an F-measurable function of ω whenever x is a Lebesgue-measurable function of t and an F-measurable function of ω; and

(iv) $\psi^{(i)}(x(t,\omega), t,\omega)\varepsilon H_{(N)}(R^+,\Omega)$ whenever $x\varepsilon H_{(N)}(R^+,\Omega)$.

Also, let

$$C_1(\psi) = \{\omega\varepsilon\Omega:\text{(i) and (ii) above are true for } \psi\varepsilon\eta_{(N)}\}.$$

We will also require the following concept.

Definition 1. A stochastic N×N matrix $A(\omega)$, whose elements are measurable functions of ω, is said to be *stochastically stable* if for some $\gamma>0$,

$$P\{\omega:\mathrm{Re}(\lambda_k(\omega)) < -\gamma,\ k=1,\cdots,\ N\} = 1,$$

where $\lambda_k(\omega)$, $k=1,\cdots, N$, are the eigenvalues of $A(\omega)$.

III. MAIN RESULTS

The systems which we consider may be viewed as interconnections of m lower order and simpler subsystems. The output of the i^{th} subsystem is given by $x_i(t,\omega)$, $x_i \varepsilon E_2$, while the input to the i^{th} subsystem is specified by $f_i(t,\omega) + \sum_{j=1}^{m} d_{ij}(\omega)x_j(t,\omega)$, where $f_i \varepsilon L_2(R^+, L_\infty(\Omega))$ and where $d_{ij}(\omega)$ represent the (random) interconnecting structure gains. The transformation from the i^{th} input to the i^{th} output will be given by the stochastic operator H_i. The specific types of operators H_i which we will consider are specified by Eqs. (1)-(3). Thus, we consider composite systems which may be represented by functional equations of the form

$$x_i(t,\omega) = H_i[f_i(t,\omega) + \sum_{k=1}^{m} d_{ik}(\omega)x_k(t,\omega)], \qquad (S)$$

$i=1,\cdots, m$. (In view of Remark 1, preceding Theorem 1, it is actually possible to consider a more general interconnecting structure in the present context.)

Subsequently, the following types of stability will be of interest.

Definition 2. System (S) is said to be *stochastically* $L_2(R^+)$ *stable* if $P\{\omega:x_i \varepsilon L_2(R^+)\} = 1$, $\{i=1,\cdots, m\} \triangleq M$.

Definition 3. System (S) is said to be *stochastically absolutely stable* if

$$P\{\omega:x_i(t,\omega) \rightarrow 0 \text{ as } t \rightarrow \infty\} = 1.$$

In characterizing the operators H_i we consider differential equations of the form

$$\left.\begin{aligned} \frac{dz_i}{dt}(t,\omega) &= A_i(\omega)z_i(t,\omega) - v_i(\omega)\psi_i(x_i(t,\omega),\ t,\ \omega) + \sum_{k=1}^{m} d_{ik}(\omega)x_k \\ x_i(t,\omega) &= f_i(t,\omega) + c_i^T(\omega)z_i(t,\omega) \end{aligned}\right\} \qquad (1)$$

$i=1,\cdots, m$, where z_i is an $N_i\times 1$ random process belonging to $E_{s(N_i)}$, $A_i(\omega)$ is a stochastically stable $N_i\times N_i$ matrix, $z_i(0,\omega)$, $c_i(\omega)$ and $v_i(\omega)$ are $N_i\times 1$ random vectors belonging to $B_{(N_i)}(\Omega)$, $\psi_i \varepsilon \eta_{(1)}$ with gain sector parameters a_i and b_i, $x_i(t,\omega)$ is a scalar random process belonging to E_2, $d_{ik} \varepsilon B_{(N_i)}(\Omega)$ are the random interconnecting structure gains, and $f_i \varepsilon L_2(R+, L_\infty(\Omega))$ denotes the i^{th} subsystem input. We impose the following additional restrictions on system (1):

(i) $|\{1 - \frac{1}{2}(a_i+b_i)c_i^T(\omega)[sI - A_i(\omega)]^{-1}v_i(\omega)\}^{-1}| \neq 0$,
$\mathrm{Re}(s) \geq 0$ a.e. [P]; and

(ii) there exist finite real numbers α_i and β_{ik} such that

$$\alpha_i \geq \sup_{\lambda \varepsilon R}|\{1 - \tfrac{1}{2}(a_i+b_i)c_i^T(\omega)[j\lambda I-A_i(\omega)]^{-1}v_i(\omega)\}^{-1} \times \{c_i^T(\omega)[j\lambda I-A_i(\omega)]^{-1}v_i(\omega)\}| \quad \text{a.e. [P], and}$$

$$\beta_{ik} \geq \sup_{\lambda\varepsilon R} \left| \{1 - \tfrac{1}{2}(a_i+b_i) c_i^T(\omega)[j\lambda I - A_i(\omega)]^{-1} v_i(\omega)\}^{-1} \times \{c_i^T(\omega)[j\lambda I - A_i(\omega)]^{-1} d_{ik}(\omega)\} \right| \quad \text{a.e. } [P].$$

If composite system (S) is characterized by Eq. (1), we define $\tilde{K}_i(s,\omega)$ by

$$\tilde{K}_i(s,\omega) = c_i^T(\omega)[sI - A_i(\omega)]^{-1} v_i(\omega).$$

In characterizing H_i of (S) we also consider stochastic differential equations of the form

$$\left.\begin{aligned} \frac{dz_i}{dt}(t,\omega) &= A_i(\omega) z_i(t,\omega) - v_i(\omega)\psi_i(x_i(t,\omega),\ t,\omega) + \sum_{k=1}^{m} d_{ik}(\omega) x_k(t,\omega) \\ x_i(t,\omega) &= f_i(t,\omega) + \int_0^t c_i^T(t-\tau,\omega) z_i(\tau,\omega)\, d\tau, \end{aligned}\right\} \quad (2)$$

$i=1,\cdots,m$, where z_i, $A_i(\omega)$, $z_i(0,\omega)$, $v_i(\omega)$, ψ_i, f_i, x_i and d_{ik} are defined as in Eq. (1) and $c_i \varepsilon L_1(R^+, L_\infty(\Omega))$ is an $N_i\times 1$ weighting function and

$$\tilde{c}_i(s,\omega) = \int_0^\infty c_i(t,\omega) e^{-st} dt.$$

We impose the following additional restrictions on Eq. (2):

(i) $|\tilde{c}_i^T(s,\omega)[sI - A_i(\omega)]^{-1} v_i(\omega)| \neq 0$ for $\mathrm{Re}(s) \geq 0$ a.e. [P];

(ii) there exist real numbers α_i and β_{ik} such that

$$\alpha_i \geq \sup_{\lambda\varepsilon R} \left| \{1 - \tfrac{1}{2}(a_i+b_i) \tilde{c}_i^T(j\lambda,\omega)[j\lambda I - A_i(\omega)]^{-1} v_i(\omega)\}^{-1} \times \{\tilde{c}_i^T(j\lambda,\omega)[j\lambda I - A_i(\omega)]^{-1} v_i(\omega)\} \right| \quad \text{a.e. } [P], \text{ and}$$

$$\beta_{ik} \geq \sup_{\lambda\varepsilon R} \left| \{1 - \tfrac{1}{2}(a_i+b_i) \tilde{c}_i^T(j\lambda,\omega)[j\lambda I - A_i(\omega)]^{-1} v_i(\omega)\}^{-1} \times \{\tilde{c}_i^T(j\lambda,\omega)[j\lambda I - A_i(\omega)]^{-1} d_{ik}(\omega)\} \right| \quad \text{a.e. } [P].$$

In connection with Eq. (2) we also use the notation

$$\tilde{K}_i(s,\omega) = \tilde{c}_i^T(s,\omega)[sI - A_i(\omega)]^{-1} v_i(\omega).$$

Finally, in characterizing H_i of (S) we also consider stochastic differential equations given by

$$\left.\begin{aligned} \frac{dz_i}{dt}(t,\omega) &= A_i(\omega) z_i(t,\omega) - \int_0^t b_i(t-s,\omega)\psi_i(x_i(s,\omega),\ s,\ \omega)\, ds \\ &\quad + \sum_{k=1}^{m} d_{ik}(\omega) x_k(t,\omega) \\ x_i(t,\omega) &= f_i(t,\omega) + \int_0^t c_i^T(t-\tau,\omega) z_i(\tau,\omega)\, d\tau \end{aligned}\right\} \quad (3)$$

where $z_i(t,\omega)$ is an $N_i\times 1$ random process belonging to $E_{s(N_i)}$, $A_i(\omega)$ is an $N_i\times N_i$ stochastically stable matrix, $z_i(0,\omega)\varepsilon B_{(N_i)}(\Omega)$, b_i, $c_i \varepsilon L_1(R^+, L_\infty(\Omega))$ are $N_i\times 1$ random weighting functions with Laplace transforms $\tilde{b}_i(s,\omega)$ and $\tilde{c}_i(s,\omega)$, respectively, $\psi_i \varepsilon \eta_{(1)}$ with gain sector parameters a_i and b_i, $f_i \varepsilon L_2(R^+, L_\infty(\Omega))$ is the input to the i^{th} subsystem, and $d_{ik}\varepsilon X_{(N_i)}$ are the random interconnecting structure gains.

We impose the following additional restrictions on system (3):

(i) $|1 - \tfrac{1}{2}(a_i+b_i)\tilde{c}_i^T(s,\omega)[sI - A_i(\omega)]^{-1} b_i(s,\omega)| \neq 0$ for $\mathrm{Re}(s) \geq 0$, a.e. [P];

(ii) there exist finite real numbers α_i and β_{ik} such that

$$\alpha_i \geq \sup_{\lambda \varepsilon R}\Big|\{1 - \tfrac{1}{2}(a_i+b_i)\tilde{c}_i^T(j\lambda,\omega)[j\lambda I-A_i(\omega)]^{-1}\tilde{b}_i(j\lambda,\omega)\}^{-1}$$

$$\times\{\tilde{c}_i^T(j\lambda,\omega)[j\lambda I-A_i(\omega)]^{-1}\tilde{b}_i(j\lambda,\omega)\}\Big| \quad \text{a.e. [P], and}$$

$$\beta_{ik} \geq \sup_{\lambda \varepsilon R}\Big|\{1 - \tfrac{1}{2}(a_i+b_i)\tilde{c}_i^T(j\lambda,\omega)[j\lambda I-A_i(\omega)]^{-1}\tilde{b}_i(j\lambda,\omega)\}^{-1}$$

$$\times\{\tilde{c}_i^T(j\lambda,\omega)[j\lambda I-A_i(\omega)]^{-1}d_{ik}(\omega)\}\Big| \quad \text{a.e. [P].}$$

For. Eq. (3) we also define $\tilde{K}_i(s,\omega)$ by

$$\tilde{K}_i(s,\omega) = \tilde{c}_i^T(s,\omega)[sI-A_i(\omega)]^{-1}\tilde{b}_i(s,\omega).$$

Remark 1. In the preceding, we can more generally consider any of the interconnecting terms d_{ik} as linear stochastic convolution operators of the form

$$d_{ik}x_k(t,\omega) = \int_0^t \bar{d}_{ik}(t-\tau,\omega)x_k(\tau,\omega)d\tau,$$

where $\bar{d}_{ik}\varepsilon L_1(R^+, L_\infty(\Omega))$. In computing the parameters β_{ik}, we merely replace $d_{ik}(\omega)$ by $\tilde{d}_{ik}(j\lambda,\omega)$ in the inequality defining the appropriate β_{ik}, where $\tilde{d}_{ik}(s,\omega)$ denotes the Laplace transform of $\bar{d}_{ik}(t,\omega)$.

We are now in a position to present the main results of this paper.

Theorem 1. Composite system (S) with subsystems described by Eqs. (1)-(3) is stochastically L_2-stable (see Definition 2) if the matrix $C=[c_{ik}]$ has positive successive principal minors, where

$$c_{ik} = \begin{cases} 1 - \alpha_i - \beta_{ii}, & i=k \\ -\beta_{ik} & , i\neq k. \end{cases}$$

Corollary 1. Composite system (S) with subsystems described by Eqs. (1)-(3) is stochastically absolutely stable if the matrix C of Theorem 1 has positive successive principal minors and if

$$P\{f_i(t,\omega) \rightarrow 0 \text{ as } t \rightarrow \infty\} = 1,\ i=1,\cdots, m.$$

Remark 2. The parameters α_i and β_{ik} may be determined graphically as follows. Define

$$\Delta_i = \frac{b_i-a_i}{2\alpha_i}\ ,\ \sigma_i = \frac{a_i+b_i}{2}\ ,\ \rho_i = \sigma_i^2-\Delta_i^2\ ,$$

and consider the following two cases:

(a) If $\rho_i>0$, $0<\alpha_i<1$, we seek the smallest $\alpha_i>0$ such that the locus of $\tilde{K}_i(j\lambda,\omega)$, $\lambda\varepsilon R$, does not touch nor encircle a circle with center $-\sigma_i/\rho_i$ and radius $|\Delta_i/\rho_i|$, a.e. [P].

(b) If $\rho_i<0$, $0<\alpha_i<1$, we seek the smallest $\alpha_i>0$ such that the locus of $\tilde{K}_i(j\lambda,\omega)$, $\lambda\varepsilon R$, is contained in a circle with center $-\sigma_i/\rho_i$ and radius $|\Delta_i/\rho_i|$ a.e. [P].

If ρ_i changes sign for $\alpha_i\varepsilon(0,1)$, an iterative process must be used to determine the smallest α_i such that the encirclement condition of either (a) or (b) is met with the appropriate sign for ρ_i. Thus, if $\rho_i>0$ we require that the locus of $\tilde{K}_i(j\lambda,\omega)$ avoids a circle with center $-\sigma_i/\rho_i$ and radius $|\Delta_i/\rho_i|$ with probability one (w.p.1.) and if $\rho_i<0$ we require that $\tilde{K}_i(j\lambda,\omega)$

be contained in a circle with center $-\sigma_i/\rho_i$ and radius $|\Delta_i/\rho_i|$ w. p. 1.

The derivation of the above circle conditions is given in [6].

IV. AN EXAMPLE

Consider a system which can be represented by the following set of stochastic differential equations:

$$\dot{z}_1(t,\omega) = A_1(\omega)z_1(t,\omega)-\psi_1(x_1(t,\omega),\ t,\ \omega)+.89d_{12}(\omega)x_2(t,\omega)$$

$$x_1(t,\omega) = 0.1\ z_1(t,\omega)+f_1(t,\omega);$$

$$\dot{z}_2(t,\omega) = \begin{bmatrix} 0 & 1 \\ -3 & -4 \end{bmatrix} z_2(t,\omega) - \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \psi_2(x_2(t,\omega),t,\omega)+d_{23}(\omega)\begin{bmatrix} .37 \\ .37 \end{bmatrix} x_3(t,\omega)$$

$$x_2(t,\omega) = f_2(t,\omega) + \int_0^t \begin{bmatrix} e^{-2(t-\tau)} \\ e^{-2(t-\tau)} \end{bmatrix} z_2(\tau,\omega)d\tau\ ;$$

$$\dot{z}_3(t,\omega) = -5.0\ z_3(t,\omega) - \int_0^t e^{-3(t-\tau)}\psi_3(x_3(\tau,\omega),\ \tau,\ \omega)d\tau$$

$$+\ 0.3\ x_1(t,\omega) + 0.7\ x_2(t,\omega)$$

$$x_3(t,\omega) = f_3(t,\omega) + \int_0^t e^{-3(t-\tau)}z_3(\tau,\omega)d\tau.$$

Assume that $-3.5\leq A_1(\omega)\leq -.73$ a.e. [P], and that $\psi_i\varepsilon\eta_{(1)}$, i=1,2,3, with sector parameters $a_1=0$, $b_1=1$, $a_2=-0.1$, $b_2=7.86$, $a_3=0$, $b_3=84.12$. We will determine a range in which $|d_{12}(\omega)|$ and $|d_{23}(\omega)|$ must lie w.p.1 to assure stochastic L_2-stability for this system.

Straightforward computations yield $\alpha_1=0.71$, $\alpha_2=0.42$, $\alpha_3=0.34$, $\beta_{12}=0.63\bar{d}_{12}$, $\beta_{23}=0.25\bar{d}_{23}$, $\beta_{31}=0.30$, $\beta_{32}=0.70$, and $\beta_{13}=\beta_{21}=0$, where $\bar{d}_{12}=\operatorname{ess\,sup}_{\omega\varepsilon\Omega}|d_{12}(\omega)|$ and $\bar{d}_{23}=\operatorname{ess\,sup}_{\omega\varepsilon\Omega}|d_{23}(\omega)|$. The test matrix C of Theorem 1 is now given by

$$C = \begin{bmatrix} 0.29 & -0.63\bar{d}_{12} & 0 \\ 0 & 0.58 & -0.25\bar{d}_{23} \\ -0.30 & -0.70 & 0.66 \end{bmatrix}.$$

This matrix has positive successive principal minors if and only if

$$2.35 - 1.075\bar{d}_{23} - \bar{d}_{12}\bar{d}_{23} > 0. \tag{4}$$

Thus, if $|d_{12}(\omega)|$ and $|d_{23}(\omega)|$ fall in the region determined by (4) w.p.1., then according to Theorem 1, the above system will be stochastically L_2-stable.

REFERENCES

[1] Porter, D. W., and Michel, A. N. "Input-Output Stability of Time-Varying Nonlinear Multiloop Feedback Systems." IEEE Transactions Automatic Control 19 (August, 1974): 422-427.

[2] Lasley, E. L., and Michel, A. N. "Input-Output Stability of Interconnected Systems." IEEE Transactions Automatic Control 21 (February, 1976): 84-89.

[3] Lasley, E. L., and Michel, A. N. "L_∞- and ℓ_∞-Stability of Intercon-

nected Systems." *IEEE Transactions Circuits and Systems* 23 (May, 1976): 261-270.

[4] Miller, R. K., and Michel, A. N. "L_2-Stability and Instability of Large Scale Systems Described by Integrodifferential Equations." (To appear) *SIAM Journal Math. Analysis*.

[5] Miller, R. K., and Michel, A. N. "Stability of Multivariable Feedback Systems Containing Elements Which Are Open-Loop Unstable." *Proceedings of the 1975 Allerton Conference on Circuit and System Theory*, Urbana, Illinois (October, 1975): 580-589.

[6] R. L. Gutmann and A. N. Michel, "Stability of Interconnected Systems Described by Stochastic Nonlinear Volterra Integral Equations", to appear.

[7] Tsokos, C. P. "The Method of V. M. Popov for Differential Systems with Random Parameters". *Journal of Applied Probability* 8 (1971): 289-310.

[8] Tsokos, C. P., and Padgett, W. J. *Random Integral Equations with Applications to Life Sciences and Engineering*. New York: Academic Press, 1974.

APPENDIX

The proofs of Theorem 1 and Corollary 1 are based on two results (Lemma A1 and Lemma A2) for composite systems described by functional equations of the form

$$\left.\begin{aligned} e_i(t,\omega) &= u_i(t,\omega) - \int_0^t k_i(t-\tau,\omega)\psi_i(e_i(\tau,\omega),\ \tau,\ \omega)d\tau \\ u_i(t,\omega) &= r_i(t,\omega) + \sum_{k=1}^{m} D_{ik}e_k \end{aligned}\right\} \quad (A1)$$

where $e_i \varepsilon E_{2(N_i)}$, $\psi_i \varepsilon \eta(N_i)$ with sector parameters a_i and b_i, $k_i \varepsilon L_1(R^+, L_\infty(\Omega)) \cap L_2(R^+, L_\infty(\Omega))$, $r_i \varepsilon L_{2(N_i)}(R^+, L_\infty(\Omega))$, and D_{ik} are operators on $E_{2(N_k)}$ with range in $E_{2(N_i)}$. For the proof of Lemma A1 and Lemma A2, refer to [6].

Lemma A1. System (A1) is *stochastically* L_2-*stable* if

(i) $\det [I + \frac{1}{2}(a_i+b_i)\tilde{K}_i(s,\omega)] \neq 0$ for $Re(s) \geq 0$, a.e. $[P]$, $\{i=1,\cdots,m\} \overset{\Delta}{=} M$, where $\tilde{K}_i(s,\omega)$ is the Laplace transform of $k_i(t,\omega)$; and

(ii) the test matrix $C=[c_{ij}]$ has positive successive principal minors, where

$$c_{ik} = \begin{cases} 1 - \alpha_i - \beta_{ii}, & i=k \\ -\beta_{ik}, & i \neq k,\ i,k \varepsilon M, \end{cases}$$

where

$$\alpha_i \geq \sup_{\lambda \varepsilon R} \Lambda\{[I + \tfrac{1}{2}(a_i+b_i)\tilde{K}_i(j\lambda,\omega)]^{-1}K_i(j\lambda,\omega)\} \quad \text{a.e. } [P],$$

and

$$\beta_{ik} \geq \|(I + \tfrac{1}{2}(a_i+b_i)K_i^{-1}D_{ik}\| \quad \text{a.e. } [P],$$

$i,k \varepsilon M$, where $\|\cdot\|$ denotes the operator norm on $L_2(R^+, L_\infty(\Omega))$.

Lemma A2. System (A1) is stochastically absolutely stable if (i) and (ii) of Lemma A1 hold and if $r_i(t,\omega)\to 0$ as $t\to\infty$.

Proof of Theorem 1. Consider first Eq. (1). Using standard differential equation techniques we have

$$z_i(t,\omega) = e^{A_i(\omega)t} z_i(0,\omega) - \int_0^t e^{A_i(\omega)(t-\tau)} v_i(\omega)\Psi_i(x_i(\tau,\omega),\tau,\omega)d\tau$$

$$+ \sum_{k=1}^{m} \int_0^t e^{A_i(\omega)(t-\tau)} d_{ik}(\omega)x_j(\tau,\omega)d\tau$$

with

$$x_i(t,\omega) = c_i^T(\omega)z_i(t,\omega) + f_i(t,\omega).$$

Then

$$x_i(t,\omega) = c_i^T(\omega)e^{A_i(\omega)t} z_i(0,\omega) - \int_0^t c_i^T(\omega)e^{A_i(\omega)(t-\tau)} v_i(\omega)\Psi_i(x_i(\tau,\omega),\tau,\omega)d\tau$$

$$+ \int_0^t c_i^T(\omega)e^{A_i(\omega)(t-\tau)} f_i(\tau,\omega)d\tau + \sum_{k=1}^{m} \int_0^t c_i^T(\omega)e^{A_i(\omega)(t-\tau)} d_{ik}(\omega)x_k(\tau,\omega)d\tau,$$

which is precisely of the form of Eq. (A1), where

$$e_i(t,\omega) = x_i(t,\omega) \qquad r_i(t,\omega) = e^{A_i(\omega)t} z_i(0,\omega) + f_i(t,\omega)$$

$$k_i(t,\omega) = c_i^T(\omega)e^{A_i(\omega)t} v_i(\omega), \text{ and}$$

$$D_{ik}e_k(t,\omega) = \int_0^t c_i^T(\omega)e^{A_i(\omega)(t-\tau)} d_{ik}(\omega)e_k(\tau,\omega)d\tau.$$

Next, we consider Eq. (2). We have

$$z_i(t,\omega) = e^{A_i(\omega)t} z_i(0,\omega) - \int_0^t e^{A_i(\omega)(t-\tau)} v_i(\omega)\Psi_i(x_i(\tau,\omega),\tau,\omega)d\tau$$

$$+ \sum_{k=1}^{m} \int_0^t e^{A_i(\omega)(t-\tau)} d_{ik}(\omega)x_k(\tau,\omega)d\tau.$$

Then

$$x_i(t,\omega) = f_i(t,\omega) + \int_0^t c_i^T(t-\tau,\omega)[e^{A_i(\omega)\tau} z_i(0,\omega)$$

$$+ \int_0^\tau e^{A_i(\omega)(\tau-s)} v_i(\omega)\Psi_i(x_i(s,\omega),s,\omega)ds$$

$$+ \sum_{k=1}^{m} \int_0^\tau e^{A_i(\omega)(\tau-s)} d_{ik}(\omega)x_k(s,\omega)ds]d\tau.$$

Consider now the integral

$$\int_0^t c_i^T(t-\tau,\omega)\int_0^\tau e^{A_i(\omega)(\tau-s)} v_i(\omega)\Psi_i(x_i(s,\omega),s,\omega)ds\, d\tau$$

$$= \int_0^t c_i^T(\tau,\omega)\int_0^{t-T} e^{A_i(\omega)(t-\tau-s)} v_i(\omega)\Psi_i(x_i(s,\omega),s,\omega)ds\, d\tau$$

$$= \int_0^t[\int_0^{t-s} c_i^T(\tau,\omega)e^{A_i(\omega)(t-s-\tau)} v_i(\omega)d\tau]\Psi_i(x_i(s,\omega),s,\omega)ds,$$

which is of the form $\int_0^t k_i(t-s,\omega)\Psi_i(x_i(s,\omega),s,\omega)ds$, where $k_i(t)= c_i(t,\omega)*e^{A_i(\omega)t}v_i(\omega)$. Hence, Eq. (2) reduces to Eq. (A1) with k_i given as above and $e_i(t,\omega)=x_i(t,\omega)$, $\Psi_i=\Psi_i$,

$$r_i(t,\omega) = f_i(t,\omega) + \int_0^t c_i^T(t-\tau,\omega)e^{A_i(\omega)\tau}z_i(0,\omega)d\tau,$$ and

$$D_{ik}e_k(t,\omega) = \int\bar{d}_{ik}(t-\tau,\omega)e_k(\tau,\omega)d\tau,$$ where

$\bar{d}_{ik}(t)=c_i^T(t)*e^{A_i(\omega)t}d_{ik}(\omega)$.

Next, consider Eq. (3). We have

$$z_i(t,\omega) = e^{A_i(\omega)t}z_i(0,\omega)$$
$$-\int_0^t e^{A_i(\omega)(t-\tau)}[\int_0^\tau b_i(\tau-s,\omega)\Psi_i(x_i(s,\omega),s,\omega)ds]d\tau$$
$$+ \sum_{k=1}^m \int_0^t e^{A_i(\omega)(t-\tau)}d_{ik}(\omega)x_k(\tau,\omega)d\tau$$

so that

$$x_i(t,\omega) = f_i(t,\omega) + \int_0^t c_i^T(t-u,\omega)\{e^{A_i(\omega)u}z_i(0,\omega) - \int_0^u e^{A_i(\omega)(u-\tau)}$$
$$\times [\int_0^\tau b_i(\tau-s,\omega)\Psi_i(x_i(s,\omega),s,\omega)ds]d\tau + \sum_{k=1}^m \int_0^u e^{A_i(\omega)(u-\tau)}d_{ik}(\omega)x_k(\tau,\omega)d\tau\}du.$$

Consider the integral

$$\int_0^u e^{A_i(\omega)(u-\tau)}[\int_0^\tau b_i(\tau-s,\omega)\Psi_i(x_i(s,\omega),s,\omega)ds]d\tau$$
$$= \int_0^u k_{1i}(u-s,\omega)\Psi_i(x_i(s,\omega),s,\omega)ds$$

where $k_{1i}(t,\omega)=e^{A_i(\omega)t}*b_i(t,\omega)$. Then

$$\int_0^t c_i^T(t-u,\omega)\{\int_0^u k_{1i}(u-s,\omega)\Psi_{i1}(x_i(s,\omega),s,\omega)ds\}du$$
$$= \int_0^t k_i(t-s,\omega)\Psi_i(x_i(s,\omega),s,\omega)ds$$

where

$$k_i(t,\omega) = c_i^T(t,\omega)*k_{i1}(t,\omega) = c_i^T(t,\omega)*e^{A_i(\omega)t}*b_i(t,\omega).$$

Hence, Eq. (3) reduces to Eq. (A1) with $k_i(t,\omega)$ as given above, with $e_i(t,\omega)=x_i(t,\omega)$, $\Psi_i=\Psi_i$,

$$r_i(t,\omega) = f_i(t,\omega) + \int_0^t c_i^T(t-u,\omega)e^{A_i(\omega)u}z_i(0,\omega)du$$

and

$$D_{ik}e_k(t,\omega) = \int_0^t \bar{d}_{ik}(t-\tau,\omega)x_k(\tau,\omega)d\tau,$$

where $\bar{d}_{ik}(t,\omega) = c_i^T(t,\omega)*e^{A_i(\omega)t}d_{ik}(\omega)$.

Therefore, the conditions imposed on Eqs. (1)-(3) and the conditions of Theorem 1 imply that all conditions of Eq. (A1) and Lemma A1 are satisfied. In addition, the conditions of Corollary 1 guarantee that all conditions of Lemma A2 are satisfied. Hence, Theorem 1 and Corollary 1 follow.

AN INTEGRATED CIRCUIT FAULT MODEL FOR DIGITAL SYSTEMS*

MARK KETELSEN
Electrical Engineering Department
Northwestern University
Evanston, Illinois 60201

ABSTRACT

A new approach to the modeling of failures in digital systems is presented. This model, the pin fault assumption, presumes an integrated circuit implementation of the system, and is based on the dominant failure mechanisms observed in such systems. Fault detection techniques based on this assumption are developed for combinational and sequential circuits.

1. Introduction

Reliability has been recognized as a primary engineering consideration in the development of digital systems for quite some time. A variety of approaches to the achievement of "fault-tolerant" digital systems have been actively researched in recent years [1,2,3]. Central to this research is knowledge of the behavior of digital circuits which have suffered physical malfunctions. Based on this knowledge, a model for faults in such systems may be defined. This model should be defined in such a way that faulty circuit analysis is simplified to the greatest extent possible within the limits of adequate description of the physical failure modes expected ("fault coverage").

Most previous research on fault-tolerant digital systems has been based on the "stuck-line"fault model [1,2], where it is assumed that faulty circuits behave as if some gate input or output were permanently "stuck" at a logical constant (0 or 1). This (single fault) model is effective in describing faulty circuit behavior only in situations where the circuit is known to be fault-free at some point in time and is subsequently diagnosed sufficiently often to assure a negligible probability that more than one failure has occurred. In other situations, it is necessary to relax the constraint that a single line is stuck, and consider multiple faults where some subset of the gate inputs and outputs of the circuit are stuck at logical 1's and 0's. This "multiple-fault" assumption has the further advantage that many failures such as certain bridging failures, shorts, and induced faults which are not modeled under the single fault assumption are equivalent to multiple stuck-line faults. The stuck-line fault model under either the single or multiple fault assumption facilitates analysis of faulty circuit behavior because the occurrence of a fault results in a well-defined logical transformation of the circuit.

The development of integrated circuit technology has resulted in widespread use of medium-and-large-scale integrated circuits in digital system implementation. The enormous increase in logical complexity afforded by this technology has resulted in a serious dilemma with respect to fault detection and related problems: Existing techniques which are based on the stuck-line fault model are impractical from a computational standpoint when applied to systems of the complexity typical of modern digital designs. In

---

* This work was supported by the Joint Services Electronics Program (U.S. Army, U.S. Navy, and U.S. Air Force) under Contract DAAB-07-72-C-0259.

such systems, the single fault assumption is invalid, while the multiple fault assumption requires impractical amounts of computation (the number of multiple faults increases at a rate which is exponential in the number of gate inputs and outputs [5]). One alternative is to resort to probabilistic testing methods [4,5].

This paper proposes a different alternative based on a new fault model, the "pin fault" model, which appears to retain acceptable fault coverage while allowing significant reductions in the computation required for generation of diagnostic tests. The rationale for this fault model is given, and fault detection techniques based on the model are presented.

2. The Pin Fault Model

The fault model proposed here is similar to the stuck-line model described above in that faults are represented as lines in the circuit stuck at logical constants, however such stuck-at faults are assumed to occur only on lines which correspond to input and output pins of the integrated circuits (IC's) used to implement the circuit. A pin fault F_i consists of one or more such stuck-line components $\{f_1^{\alpha}, f_2^{\beta}, \cdots\}$, where f_1^{α} denotes the fault component "pin 1 stuck-at-α." (It is assumed that all pin faults are "proper" faults, i.e. no pin fault contains both f_k^{0} and f_k^{1} for any pin k.)

The pin fault model is dependent on the specific IC implementation of a given logic circuit. The stuck-pin components of the pin fault model represent a subset of the stuck-line components of the stuck-line fault model. This reduction in the number of potential fault sites is most dramatic when IC's with a high ratio of internal lines to pins are present. Such high ratios are typical of IC technology, where the number of external pins required is often the limiting factor in determining the amount of logic which can be placed on an IC chip. This reduction in the number of potential fault sites, in conjunction with the implicit assumption that all logic within IC boundaries is fault-free has the potential for dramatic simplification of the fault detection problem. The cost of this simplification is a potential reduction in fault coverage.

Available failure data for several families of digital IC's show the dominant failure mode to be "lead-bond" failure, i.e. failures occurring at the interconnection between the IC package pins and the silicon chip itself [6,7]. Such failures result from thermal and mechanical shock. Failures resulting from electrical overload typically result in faulty input or output transistors. Overall system reliability depends upon the reliability of the interconnections between IC's in addition to the reliability of the IC's themselves. Open solder joints and printed circuit board routing have been identified as dominant system failure modes [7]. The pin fault model adequately describes all of these failure modes, and therefore adequate fault coverage is provided by this model when these modes (and others corresponding to pin fault failures) represent the dominant causes of failures.

3. Combinational Circuits

The most basic context for the consideration of fault detection is at the level of the single output IC module. Let $Z(X,F_i)$ denote the output of an ℓ-input module under input $X = x_1x_2\ldots x_\ell$ when pin fault F_i is present, and let $Z(X,\phi)$ be the fault-free output. $T = \{t_1,t_2,\cdots\}$ constitutes a pin fault detection test set for such a module if for every pin fault F_i, there is some $t_k \in T$ such that $Z(t_k,F_i) \neq Z(t_k,\phi)$, where t_k is an ℓ-element input vector. The problem to be considered in this section is the determination of such a test set.

Pin fault component f_j^α is detected completely by partial test set $\tau \subset T$ if and only if for every pin fault F_i which contains f_j^α as a component there is a test, $t_k \in \tau$, such that $Z(t_k,F_i) \neq Z(t_k,\phi)$. A test set may be derived by deducing a set of input vectors which detect completely all pin fault components. (This strategy allows fault components to be individually considered rather than in combination with other components,) The following theorem identifies a class of test sets with the property that each pin i (carrying input x_i[*]), there is a pair of tests which completely detect the pin fault components f_i^0 and f_i^1:

Theorem 1: $T = \{t_1,t_2,\cdots\}$ is a test set for an IC module realizing function $Z(X,\phi)$ if for every input x_i there exists a pair of test vectors t_j and t_k in T such that t_j and t_k differ only in the value applied to x_i and $Z(t_j,\phi) \neq Z(t_k,\phi)$.

Theorem 2: Any ℓ-input IC module which realizes a nonvacuous function of the ℓ input variables possesses a test set satisfying Theorem 1.

The generation of test sets of the type defined above may proceed from a functional specification of the IC without regard for the internal realization. It is assumed that such a specification in terms of a covering set of prime implicants is available.

Consider a prime implicant P_k for ℓ-input function $Z(X,\phi)$ given in "cube notation" where the symbol "u" represents an unspecified variable, and subscripted greek letters represent an assignment of values to the specified variables. (The variables are permuted so that unspecified variables of the prime implicants appear in the first r variable positions.)

$$P_k = uu \cdots u\alpha_{r+1} \cdots \alpha_\ell$$

P_k covers 2^r minterms of $Z(X,\phi)$. The r-cube which is adjacent to P_k in variable x_i is obtained from P_k as follows:

$$P_k^{\tilde{i}} = uu \cdots u\alpha_{r+1} \cdots \alpha_{i-1}\bar{\alpha}_i\alpha_{i+1} \cdots \alpha_\ell$$

The fact that cube P_k is a prime implicant assures that at least one vertex of $P_k^{\tilde{i}}$ is a false vertex. Then:

$$v \in P_k^{\tilde{i}} \cap \bar{Z}(X,\phi)$$

$$v = \beta_1\beta_2 \cdots \beta_r\alpha_{r+1} \cdots \bar{\alpha}_i \cdots \alpha_\ell$$

Then the following pair of vertices satisfy the hypothesis of Theorem 1 for input x_i and therefore completely detect fault components f_i^0 and f_i^1:

$$(\beta_1\beta_2 \cdots \beta_r\alpha_{r+1} \cdots \alpha_i \cdots \alpha_\ell,\ \beta_1\beta_2 \cdots \beta_r\alpha_{r+1} \cdots \bar{\alpha}_i \cdots \alpha_\ell)$$

---

[*] Since it will be unnecessary to refer explicity to fault components on pins other than those carrying input signals, it is convenient to make a correspondence between the indices of inputs and pins.

Example:

$$Z(X,\phi) = \sum (P_1,P_2,P_3)$$

$$= \sum (u100,01u1,1ulu)$$

generate test pair for input x_1 using P_2:

$$P_2 = 01u1$$

$$P_2^{\tilde{i}} = 11u1$$

$$P_2^{\tilde{i}} \cap Z(X,\phi) = 1101$$

$$\therefore \text{ tests for } x_1: \quad 1101 \in \overline{Z}(X,\phi)$$
$$0101 \in Z(X,\phi)$$

Automation of this procedure to produce a test set involves determination of a subset of the prime implicants sufficient to derive test pairs for each variable, formation of the adjacent cubes $P_k^{\tilde{i}}$, and the determination of assignments to $\beta_1 \cdots \beta_r$ which produce false vertices. This process is shown to require significantly less computation than the generation of single fault test sets under the stuck-line fault model [8].

Test sets generated by this method contain 2ℓ test vectors (for an ℓ-input module) regardless of the logical complexity of the function, and also have the property that ℓ of the test vertices are false vertices and ℓ are true vertices. This property simplifies the test procedure since application of test vectors in the proper order results in a strictly alternating output sequence for $Z(X,\phi)$ thus eliminating the need to store an array of fault-free circuit responses for comparison with the response of the circuit under test.

Theorem 1 may be satisfied with as few as $\ell+1$ test vectors for certain functions when test vertices may be shared by more than one test pair. Multiple pin fault test sets which do not satisfy Theorem 1 also exist, and certain functions may be tested with fewer than the lower bound of $\ell+1$ tests required by Theorem 1. A striking example of such a function is the following:

$$Z(X,\phi) = x_1x_2 \cdots x_\ell \vee \overline{x}_1\overline{x}_2 \cdots \overline{x}_\ell$$

$T = \{00 \cdots 0,\ 11 \cdots 1,\ v_0\}$ where v_0 is any false vertex. It may be shown that three tests are sufficient to detect all multiple pin faults of such functions, regardless of the number of input pins [8].

4. Synchronous Sequential Circuit Diagnosis

Previous research in fault diagnosis of synchronous circuits has proceeded along two fundamentally different approaches. The machine identification approach [9] assumes that faults may corrupt the flow table which describes the fault-free circuit in any manner which does not result in an increase in the number of states. This fault model is very broad in the types of faults which are allowed, and has the additional advantage of being independent of specific realizations of the flow table. However, design of diagnostic test sequences is extremely complex since, for an N state circuit, it is necessary to derive an input sequence capable of distinguishing between the fault-free circuit and any circuit which realizes a flow table with N or fewer states. Additionally, test sequences with this capability are normally quite lengthy. The second approach considers a particular circuit realization of the flow table and assumes the single stuck-line fault model [10]. This circuit testing approach simplifies test design, however, the drawbacks of the single-fault assumption must be accepted. Under the pin fault assumption, diagnostic test design is independent of circuit implementation provided that the combinational portion of the circuit contains no multi-module signal paths. Additionally, the generation of test sequences is enhanced because pin faults affect the sequential behavior of the circuit in a well-defined manner.

It is assumed that module pins carry the following types of signals: input variables $x_1 \cdots x_\ell$, output signals $Z_1 \cdots Z_m$, next state feedback signals, state variable signals, and clock signals. It is assumed that only the output signals are observable, and that sequential circuits are described by reduced, strongly connected flow tables. With these assumptions, the effects of output signals on circuit behavior are as follows: Faults with components involving state variables, next state variables, and clock pins result in a decrease in the number of states. Faults with components involving the output lines cause constant output signals on the faulty lines. Faults involving input variables cause certain columns in the flow table to effectively be replaced by other columns in the flow table.

The fault detection experiment to be described consists of three parts. Part 1 displays the existence of the correct number of states and therefore detects all faults involving the memory portion of the circuit. Part 2 detects all faults which involve input variables. Part 3 detects faults involving any output lines which are not exercised during parts 1 and 2.

Part 1 is the same as that used in the fault detection approaches of Hennie [9] and Kohavi [10] and consists of a series of distinguishing sequences and transfer sequences applied in such a way that N distinct responses to the distinguishing sequences are produced by the fault-free circuit (which possesses N states). In addition to verifying the correct number of states in the circuit under test, this part of the sequence also establishes the validity of the transfer sequences used.

Part 2 of the test sequence, which detects faults involving the primary input pins, consists of a series of sequences which test a set of flow table transitions sufficient to diagnosis all input faults. Design of this part consists of first determining such a set of transitions, and then forming the series of test sequences for them. Consider the flow table shown in Table 1. A faulty transition is detectable if either an erroneous state is entered or an erroneous output is produced. If a faulty state is entered, but the correct output is produced, a distinguishing sequence must be applied in order to detect the error. For example consider the transition from state C under input 10. If the fault $F = f_1^0$ is present, then the erroneous transition given in input column 00 will occur. Since the output produced by this transition is erroneously a 1, this fault will be detected immediately. For this particular transition, any fault which causes a transition other than that given by input column 10 will be detected. However the following faults cannot be detected by this transition: f_1^1, f_2^0, $f_1^1 f_2^0$. Consider the transition from state A under input 01. Any faults which cause the machine to enter input columns 11 or 10 will be detected because of erroneous output. If, however, column 00 is entered, the output is correct. This erroneous transfer can be detected, since an error in the state occurs. An application of the distinguishing sequence will detect such an error. Thus all faults other than f_1^0, f_2^1, $f_1^0 f_2^1$ will be detected by this transition. That these two transition are sufficient to detect all input faults may be demonstrated by forming the following set intersection:

$$\text{faults undetected by } C(10)\text{:} \quad \{f_1^1, f_2^0, f_1^1 f_2^0\}$$

$$\text{faults undetected by } A(01)\text{:} \quad \{f_1^0, f_2^1, f_1^0 f_2^1\}$$

$$\{f_1^1, f_2^0, f_1^1 f_2^0\} \cap \{f_1^0, f_2^1, f_1^0 f_2^1\} = \emptyset$$

In order to form part 2 of the test sequence it is necessary to bring the circuit to the initial state of each of the transitions to be tested using only transfer sequences verified in part 1 and then execute the transitions. In the case of the transition A(01), a distinguishing sequence must then be applied.

In order to complete the fault detection sequence, it is only necessary to insure that all outputs have been exercised. All outputs upon which both 0's and 1's have appeared are obviously fault-free. Those which have produced constant output throughout parts 1 and 2 must be forced to assume a complementary value. Any input sequence which tests all untested outputs in this manner is sufficient to complete the fault detection sequence. In the present example, part 3 is not required. A complete pin fault detection sequence fot the flow table of Table 1 is the following (Since C(10) is executed during part 1 it is unnecessary to repeat it in part 2):

x_1x_2	00	00	10	10	10	10	10	10	01	10	10
state	?	C	D	A	B	C	D	A	A	B	C
output			0	0	1	1	0	0	0	1	1

Part 1 ↑ C(10)checked

↑ Part 2 A(01) checked

The reduction in complexity of determining such sequences by this method is significant. The resulting test sequence is also substantially shortened compared to the machine identification and circuit testing approaches. Both of these methods require part 1 as given above, however the machine identification approach requires 41 additional inputs [9] and the circuit testing approach requires 15 additional inputs [10] compared to the 3 required above.

5. Asynchronous Sequential Circuit Diagnosis

The essential characteristic of asynchronous circuit operation is that such circuits respond directly to changes in the values applied to the inputs. Because of this behavior, such circuits are quite naturally diagnosed for pin faults. Under previously studied fault models, asynchronous circuit diagnosis presents significant problems [2]. For example, it is possible for stuck-line type faults to result in the creation of critical races where none exist for the fault-free circuit.

Consider the following input/output sequence for a 3-input, single output asynchronous circuit:

$x_1x_2x_3$	000	001	011	001	101	111	011	001
Z	1	0	0	0	1	1	0	1

It may be concluded from this sequence alone, without knowledge of the circuit realization or even the flow table that the circuit which produced this response is pin-fault free. Observe that as the input changed from 000 to 001 a change in the output occurred. Since only input x_3 was changed x_3 must be fault-free. If x_3 were stuck, the internal logic would not have sensed an input change, and would not have responded. Similar transitions may be observed for x_1 and x_2.

The ability to diagnose asynchronous circuits in such a straight-forward manner requires certain flow table properties. An input pin carrying x_i is unit transition testable (UTT) if the flow table contains two states Q_i and Q_j in input columns I_α and I_β with the following property: Q_j is direct successor of state Q_i under the single input change involving input x_i (i.e. I_α and I_β differ only in x_i), and $Z(Q_i,I_\alpha,\phi) \neq Z(Q_j,I_\beta,\phi)$ where $Z(Q,I,\phi)$ is the fault-free mapping of total states onto output vectors. The transition from state Q_i to Q_j is a test transition for input pin x_i. An asynchronous circuit is unit transition diagnosable (UTD) if all inputs are UTT.

Fault detection sequences for circuits possessing the UTD property may be designed by selecting one test transition for each input and then formulating a sequence which initializes the circuit to a known state and then causes the set of test transitions to be executed. It is unnecessary to verify the required transfer sequences prior to their use. If the circuit is implemented in such a way that some feedback lines are routed external to IC modules, then it is also necessary to verify that the number of states has not decreased. Since the inputs are already tested by the sequence just described, it is sufficient to verify the existence of one total state in each flow table row.

Example: (Refer to Table 2)

input pin	test transitions
x_1	B(01)→A(11), D(11)→B(01)
x_2	C(00)→C(01), B(01)→A(00), C(01)→C(00) D(11)→D(10), D(10)→D(11)

x_1x_2	00	01	00	10	11	01	00	10
state	A,C	C	C	D	D	B	A	B
Z	0,1	0	1	1	0	1	0	0

synchronizing sequence x_2 tested x_1 tested

The diagnosis of this machine using the machine identification approach is considered in [2] and a sequence of 28 inputs is required for diagnosis compared to the 8 inputs required here.

Any nondegenerate flow table must possess some UTT inputs, however not all asynchronous machines are UTD. Table 3 shows such a flow table. (This table was obtained from Table 2 by altering some of the output assignments.) It may be observed that x_2 is non-UTT. Test sequences for such circuits may be designed by first identifying the UTT inputs and forming a test sequence for these inputs as described above. Assuming that the circuit under test has responded correctly to this portion of the test sequence, the remainder of the sequence may utilize the information that the UTT inputs are fault-free. Faults on non-UTT inputs restrict operation of the machine to some submachine. The testing of non-UTT inputs may be achieved by verifying connectivity of the submachines which result from the presence of fault components involving untested inputs. For example in the circuit described by Table 3 where x_2 is non-UTT, the fault f_2^1 restricts operation to half of the state transition graph, and the fault f_2^0 restricts operation to the other half. To verify x_2 is fault-free it is sufficient to apply a sequence which verifies that a transition between these two submachines occurs when x_2 is changed. The assumption that the flow table is reduced assures that it is possible to establish the connectivity of these submachines. The following sequence diagnoses this circuit for input pin faults:

$$\begin{array}{cccccc} x_1x_2 & 01 & 11 & 01 & 00 & 10 \\ \text{state} & C,B & A & C & C & D \end{array}$$

$$Z \quad \underbrace{0 \quad 1}_{x_1 \text{ tested}} \quad \underbrace{0 \quad 0 \quad 0}_{x_2 \text{ tested}}$$

Testing for faults on feedback lines is accomplished in the same manner as described for the UTD case.

6. IC Modification to Enhance Testability

IC chips are typically encapsulated in standard packages, and as a result it is not uncommon for unused package pins to result. If such pins are available a simple modification to the IC device may be made to greatly simplify diagnosis of pin faults.

As discussed in Section 3, the function $Z(X,\phi) = x_1x_2 \cdots x_\ell \vee \overline{x}_1\overline{x}_2 \cdots \overline{x}_\ell$ requires only three tests for pin fault diagnosis regardless of the number of inputs. Such a function may be added to an IC, either combinational or sequential, to provide input diagnosis with only the three tests mentioned. One additional IC pin is required. The outputs of such a modified device are simply diagnosed by applying inputs which result in both states appearing on all outputs. The proposed modification is shown in Figure 1.

7. Conclusions

The conclusions which may be drawn from the results reported here are the following: The pin fault model appears to provide adequate fault coverage in some instances, although additional research of reliability data for IC's and systems is required to establish the applicability of the model. In instances where the model is adequate, a significant reductions in both the computation required to generate diagnostic tests, and the length of such tests may be achieved. The trend of increasing logical complexity of IC devices tends to enhance the advantages of the pin fault approach. The addition of a nominal amount of logic (3 gates) and the use of one IC package pin allows greatly enhanced pin fault diagnosis.

REFERENCES

1. H. Y. Chang, E. G. Manning, and G. Metze, Fault Diagnosis of Digital Systems, Wiley-Interscience, New York, 1970.

2. A. D. Friedman and P. R. Menon, Fault Detection in Digital Circuits, Prentice-Hall, New York, 1971.

3. M. A. Bruer and A. D. Friedman, Diagnosis and Reliable Design of Digital Systems, Computer Science Press, Woodland Hills, CA, 1976.

4. J. P. Hayes, "Testing Logic Circuits by Transition Counting," Proc. of the Fifth International Symp. on Fault-Tolerant Computing, pp. 215-219, Paris, 1975.

5. J. Losq, "Referenceless Random Testing," Proc. 1976 International Symp. on Fault-Tolerant Computing, pp. 108-113, Pittsburg, PA, 1976.

6. E. F. Platz, "Solid Logic Technology Computer Circuits-Billion Hour Reliability Data," Microelectronics and Reliability, Vol. 8, pp. 55-59, 1969.

7. J. C. Mc Donald, "Testing for High Reliability: A Case Study," Computer, Vol. 9, pp. 18-21, 1976.

8. M. L. Ketelsen, "An Integrated Circuit Fault Model for Digital Systems," Report R-743, Coordinated Science Lab., University of Illinois, Urbana, Illinois, 1976.

9. F. C. Hennie, "Fault Defection Experiments for Sequential Circuits," Proc. of Fifth Annual Symp. on Switching Circuit Theory and Logical Design, pp. 95-110, 1964.

10. I. Kohavi, "Fault Diagnosis of Logical Circuits," Proc. of Tenth Annual Symp. on Switching and Automata Theory, pp. 166-173, 1969.

		x_1x_2			
		00	01	11	10
	A	B,0	A,0	C,1	B,1
present state	B	C,0	D,0	C,1	C,1
	C	C,1	D,1	D,1	D,0
	D	C,1	D,1	A,1	A,0

next state, Z

Table 1. Synchronous flow table

	x_1x_2			
state	00	01	11	10
A	Ⓐ,0	C	Ⓐ,0	B
B	A	Ⓑ,1	A	Ⓑ,0
C	Ⓒ,1	Ⓒ,0	A	D
D	C	B	Ⓓ,0	Ⓓ,1

Table 2. UTD asynchronous flow table

	x_1x_2			
state	00	01	11	10
A	Ⓐ,0	C	Ⓐ,1	B
B	A	Ⓑ,0	A	Ⓑ,1
C	Ⓒ,0	Ⓒ,0	A	D
D	C	B	Ⓓ,0	Ⓓ,0

Table 3. Non-UTD asynchronous flow table

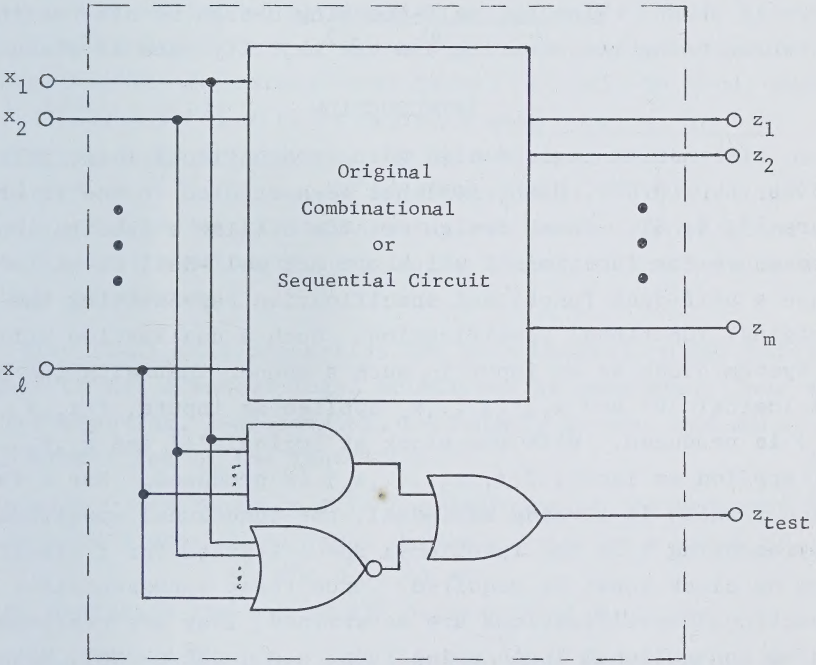

x_1
x_2
x_ℓ
Original Combinational or Sequential Circuit
z_1
z_2
z_m
z_{test}

Figure 1. IC modification for diagnosis.

SELF-CHECKING DESIGN USING COMPLETE SETS OF ALTERNATING PRIMITIVES

DENNIS A. REYNOLDS
Sandia Laboratories
Albuquerque, New Mexico

ABSTRACT

Alternating combinational design with sets of self-dual primitives is defined. Necessary and sufficient conditions that these sets be complete; that is, that any boolean function be realizable in an alternating network as a composition of these primitives, are defined. That such alternating networks are inherently self-checking for all stuck-at faults on primitive input and output lines regardless of network structure is shown. Finally, self-checking design of alternating networks using the minority and the majority gate is discussed.

INTRODUCTION

Alternating logic design using conventional logic primitives (AND/OR/NOT, NAND, NOR) has been studied in the literature [3, 4, 5]. These design methods utilize a dualization procedure for functions f which are not self-dual so as to produce a self-dual functional specification representing the original functional specification. Such a dualization utilizes a system clock as an input in such a manner that with the clock at logical '0' and $x_1,x_2,\ldots,x_n$ applied as inputs, $f(x_1,x_2,\ldots,x_n)$ is produced. With the clock at logical '1' and $\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_n$ applied as input, $\bar{f}(x_1,x_2,\ldots,x_n)$ is produced. For a function f which is already self-dual, the functional specification representing f is the functional specification for f itself, and no clock input is required. Once these representative functional specifications are determined, they are realized using conventional logic primitives, e.g., AND/OR/NOT, NAND, or NOR, in networks designed to be self-checking for all single faults. In each case, design procedures utilizing conventional sets of logic primitives lead to networks of restricted structure. These structural restrictions, however, can be completely eliminated if only alternating primitives are used as basic logic design modules in the design of alternating logic systems.

WEAK COMPLETE SETS

Consider a set F of self-dual functions $F = \{ f^{(1)}, f^{(2)}, \ldots, f^{(p)} \}$ and suppose that this set of logic primitives is to be used to realize the function f directly. That is, the function f is to be realized as a composition of functions from the set F. A composition of self-dual functions, (Ibuki [1]), is itself self-dual. Thus, while the realization of any composition of functions from the set F is an alternating circuit, only alternating circuits realizing self-dual functions result from the compositions. In alternating logic design, however, the clock is available and is itself a self-dual function representing the conventional 0 function. So, every Boolean function can be represented in an alternating circuit constructed as a composition of alternating primitives $f^{(1)}, f^{(2)}, \ldots, f^{(p)}$ if and only if $F' = \{f^{(1)}, \ldots, f^{(p)}, 0\}$ is a complete set of logic primitives in conventional terms. If this is true, then $\{f^{(1)}, f^{(2)}, \ldots, f^{(p)}\}$ will be called a weak complete set of alternating logic primitives. To determine what restrictions are placed on $f^{(1)}, f^{(2)}, \ldots, f^{(p)}$ such that $\{f^{(1)}, f^{(2)}, \ldots f^{(p)}, 0\}$ is a complete set of logic primitives in the conventional sense requires some background in the theory of complete sets of logic primitives.

Five functional properties are instrumental in determining whether or not a set of logic primitives is complete. They are the 0-preserving, 1-preserving, self-dual, linear, and monotonic properties of the functions.

Definition 1: A function f is 0-preserving if the value of the function is 0 when every input variable takes on a value of 0.

$$f(0,0,\ldots 0) = 0$$

Let M_1 designate the set of all 0-preserving functions.

Definition 2: A function f is 1-preserving if the value of the function is 1 when every input variable takes on the value 1.

$$f(1,1,\ldots,1) = 1$$

Let M_2 designate the set of all 1-preserving functions.

Definition 3: A function f is self-dual iff the function f and its dual function f^D are identical or equivalently,

$$f(x_1,x_2,\dots,x_n) = \bar{f}(\bar{x}_1,\bar{x}_2,\dots,\bar{x}_n).$$

Let M_3 designate the set of all self-dual functions.

Definition 4: A function f is linear if it can be written in the form

$$f(x_1,x_2,\dots,x_n) = g_o \oplus (g_1x_1 \oplus g_2x_2 \oplus \cdots \oplus g_nx_n).$$

Let M_4 designate the set of all linear functions.

Definition 5: A function f is monotonically increasing if for every $x_{11}\ x_{12}\ \dots x_{1n} \geq x_{21}\ x_{22}\ \dots x_{2n}$ then

$$f(x_{11},x_{12},\dots,x_{1n}) \geq f(x_{21},x_{22},\dots,x_{2n}).$$

Let M_5 designate the set of all monotonically increasing functions.

These five properties will be used in the development to follow. Also used is a convenient notation for the function set generated by a set R.

Definition 6: The function set S generated by a function set $R = \{f_1,f_2,\dots,f_q\}$ is the set of functions realizable as a composition of functions from the set R and is denoted by $[R]$.

Now, consider again the set of logic primitives $F' = \{f^{(1)}, f^{(2)},\dots,f^{(p)},0\}$ where $f^{(1)},f^{(2)},\dots,f^{(p)}$ are self-dual functions. Completeness of the set F' is dependent on the properties of the self-dual functions $f^{(1)},f^{(2)},\dots,f^{(p)}$. These dependencies are defined in the following.

Lemma 1: If $F' \not\subset M_1$, then $\bar{x} \in [F']$ and $1 \in [F']$.

Proof: If $F' \not\subset M_1$, then there exists a function $f \in F'$ which is not 0-preserving. So then, either $f \in M_2 \cap \bar{M}_1$, or $f \in \bar{M}_2 \cap \bar{M}_1$. But, if f is self-dual and not 0-preserving then

$$f(0,0,\dots,0) = 1$$

$$\Rightarrow f(1,1,\dots,1) = 0,$$

and f is not 1-preserving. So, $f \notin M_2 \cap \bar{M}_1$. That is, $f \in \bar{M}_2 \cap \bar{M}_1$. But then $\bar{x} \in [F']$, since the function $\bar{x}$ can be obtained from f by tying all inputs of f together. Thus, for 0 in, $x_1 = 0$, $x_2=0,\dots,x_n=0$ and, since f is not 0-preserving, a 1 is produced as an output. For 1 in, $x_1 = 1, x_2 = 1,\dots,x_n = 1$ and a 0 is produced as f is not 1-preserving. Now, since $\bar{x}\varepsilon[F']$ and $0\varepsilon[F']$, $1\varepsilon[F']$. Q.E.D.

Further, Ibuki [1] has proved the following.

Lemma 2: If $R \not\subset M_4$ and $\bar{x}, 0, 1 \varepsilon [R]$ then R is a complete set of primitives, where R is an arbitrary set of functions not necessarily self-dual.

Based on the preceeding lemmas the restrictions on $f^{(1)}$, $f^{(2)}, \ldots, f^{(p)}$ such that any boolean function can be represented in an alternating network as a composition of $f^{(1)}$, $f^{(2)}, \ldots, f^{(p)}$ can be stated.

Theorem 1: A set $F = \{f^{(1)}, f^{(2)}, \ldots, f^{(p)}\}$ of self-dual functions is a weak complete set of alternating logic primitives if $f^{(i)}$ is not 0-preserving for some i, $1 \leq i \leq p$, and $f^{(j)}$ is nonlinear for some j, $1 \leq j \leq p$ where possibly $i = j$.

Proof: The set F is a weak complete set of alternating logic primitives if the set $F' = \{f^{(1)}, f^{(2)}, \ldots f^{(p)}, 0\}$ is a complete set in the conventional sense. And by Lemma 1 and Lemma 2, F' is complete if $F' \not\subset M_1$ and $F' \not\subset M_4$, or in other words if $f^{(1)}$ is not 0-preserving for some i, $1 \leq i \leq p$ and $f^{(j)}$ is nonlinear is nonlinear for some j, $1 \leq j \leq p$. Q.E.D.

STRONG COMPLETE SETS

In any application of the theory outlined thus far it is implicitly assumed that the (0,1) clock, the 0-function in an alternating network, may be used as an input to any variable line in a network application of any function from the set F. This means that the (0,1) clock may be input on 2 or more variable lines in an application of a given logic primitive in the composition of a desired alternating network from the set F of alternating logic primitives. This, however, may result in an undesirable loading of the clock. If application of the clock is limited to at most one input line of any logic primitive, a somewhat different theory of alternating logic primitives results.

Once again let $F = \{f^{(1)}, f^{(2)}, \ldots, f^{(p)}\}$ where $f^{(j)}$ is a function of k_j variables, $1 \leq j \leq p$, be a set of self-dual functions. The set F is a strong complete set of alternating logic primitives if an alternating network representing any Boolean function can be constructed by composing elements of F with the provision that the (0,1) clock is used as an input to at most one variable of any given alternating logic primitive. Before the problem can be restated in terms pertinent to conventional theory of complete sets, the concept of a 0-subfunction must be defined.

Definition 7: The i-th 0-subfunction $f_{0,i}$ of a function $f(x_1, x_2, \ldots, x_n)$ of n variables is the function obtained by constraining the i-th variable in $f(x_1, x_2, \ldots, x_n)$ to a 0. That is,

$$f_{0,i}(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = f(x_1, x_2, \ldots, x_{i-1}, 0, x_{i+1}, \ldots x_n).$$

For a function $f^{(j)}$, $1 \leq j \leq p$, from the set F the i-th 0-subfunction will be denoted $f^{(j)}_{0,i}$. Now, the set $F = \{f^{(1)}, f^{(2)}, \ldots, f^{(p)}\}$ will be a strong complete set of alternating logic primitives if

$$T = \left\{ f^{(1)}, f^{(2)}, \ldots, f^{(p)}, \bigcup_{m=1}^{p} \bigcup_{i=1}^{k_m} f^{(m)}_{0,i} \right\}$$

is a set of complete logic primitives in the conventional sense. Further, the completeness of T can be determined using the following theorem, Ibuki [1].

Theorem 2: A function set R is a complete set of functions if $F \not\subset M_i$ for $1 \leq i \leq 5$.

The proof of the theorem is presented in [1] and is not repeated here. Now, the number of functions in T is at most $p + \sum_{m=1}^{p} k_m$ as some of the 0-subfunctions may be identical. At any rate it appears that each function in T must be classified according to the five properties, 0-preserving, 1-preserving, self-dual, linear, and monotonic so as to determine if T is a complete set. Fortunately, this is not the case as will be shown.

Lemma 3: No 0-subfunction $f_{0,i}$ of a self-dual function $f(x_1, x_2, \ldots, x_n)$ is self-dual, $1 \leq i \leq n$.

Proof: By hypothesis f is a self-dual function of n variables. By way of contradiction assume that $f_{0,i}$ is self-dual.

Then,

$$f(x_1,x_2,..,x_{i-1},0,x_{i+1},..,x_n) = \bar{f}(\bar{x}_1,\bar{x}_2,..,\bar{x}_{i-1},0,\bar{x}_{i+1},..,\bar{x}_n).$$

But, since f is self-dual

$$f(x_1,x_2,..,x_{i-1},1,x_{i+1},..,x_n) = \bar{f}(\bar{x}_1,\bar{x}_2,..,\bar{x}_{i-1},0,\bar{x}_{i+1},..,\bar{x}_n)$$

$$\Rightarrow f(x_1,x_2,..,x_{i-1},0,x_{i+1},..,x_n) = f(x_1,x_2,..,x_{i-1},1,x_{i+1},..,x_n)$$

for every $x_1,x_2,..,x_{i-1},x_{i+1}..,x_n$ and it therefore follows that f is a function of n-1 variables, a contradiction. Q.E.D.

Lemma 4: If $f(x_1,x_2,..,x_n)$ is a linear function of n variables, then every 0-subfunction $f_{0,i}$, $1 \leq i \leq n$, is linear.

Proof: By hypothesis f is a linear function and can therefore be written in the form

$$f = g_o \oplus (g_1x_1 \oplus g_2x_2 \oplus \ldots \oplus g_nx_n)$$

where $g_o,g_1,\ldots,g_n \varepsilon \{0,1\}$. But then,

$$f_{0,i}(x_1,x_2,..,x_{i-1},x_{i+1},..,x_n) = g_o \oplus (g_1x_1 \oplus .. \oplus g_{i-1}x_{i-1} \oplus g_{i+1}x_{i+1} \oplus .. \oplus g_nx_n).$$

The function $f_{0,i}$ is therefore a linear function also. Q.E.D.

Lemma 5: If $f(x_1,x_2,..,x_n)$ is a monotonically increasing function of n variables, then every 0-subfunction, $f_{0,i}$, $1 \leq i \leq n$, is monotonically increasing.

Proof: By way of contradiction assume that $f(x_1,x_2,..,x_n)$ is monotonically increasing but that $f_{0,i}$ for some i, $1 \leq i \leq n$, is not. Then, there exists a pair of (n-1)-tuples

$$x_{1,1}x_{2,1}\ldots x_{i-1,1}x_{i+1,1}\ldots x_{n,1} \geq x_{1,2}x_{2,2}\ldots x_{i-1,2}x_{i+1,2}..x_{n,2}$$

for which

$$f_{0,i}(x_{1,1},x_{2,1}..,x_{i-1,1},x_{i+1,1},..,x_{n,1}) < f_{0,i}(x_{1,2},x_{2,2},.., x_{i-1,2},x_{i+1,2},..,x_{n,2}).$$

But, since

$$f_{0,i}(x_{1,1},x_{2,1},..,x_{i-1,1},x_{i+1,1},..,x_{n,1}) = f(x_{1,1},x_{2,1},.., x_{i-1,1},0,x_{i+1,1},..,x_{n,1})$$

and

$$f_{0,i}(x_{1,2},x_{2,2},..,x_{i-1,2},x_{i+1,2},..,x_{n,2}) = f(x_{1,2},x_{2,2},.., x_{i-1,2},0,x_{i+1,2},..,x_{n,2})$$

for

$$x_{1,1}x_{2,1}..x_{i-1,1}x_{i+1,1}..x_{n,1} \geq x_{1,2}x_{2,2}..x_{i-1,2}x_{i+1,2}..x_{n,2},$$

$$f(x_{1,1},x_{2,1},..,x_{i-1,1},0,x_{i+1,1},..,x_{n,1})<f(x_{1,2},x_{2,2},..,x_{i-1,2}, 0,x_{i+1,2},..,x_{n,2}).$$

The function f is therefore not monotonically increasing, a contradiction. Q.E.D.

The above lemmas have thus shown that the set

$$T = \left\{f^{(1)},f^{(2)},...,f^{(p)},\bigcup_{m=1}^{p}\ \bigcup_{i=1}^{k_m} f_{0,i}^{(m)}\right\}$$

always contains a function which is not self-dual. Specifically every subfunction $f_{0,i}$ is not self-dual. Further, the lemmas show that to determine if T is a subset of the set of linear functions, $T \subset M_4$, or if T is a subset of the set of monotonically increasing functions, $T \subset M_5$, only the functions $f^{(1)},f^{(2)},..,f^{(p)}$ need be examined with respect to these properties.

Theorem 3: A set of self-dual functions $F = \{f^{(1)},f^{(2)},.., f^{(p)}\}$ is a strong complete set of alternating logic primitive if $f^{(i)}$ is not 0-preserving for some i, $1 \leq i \leq p$, $f^{(j)}$ is not linear for some j, $1 \leq j \leq p$, and $f^{(k)}$ is not monotonically increasing for some k, $1 \leq k \leq p$.

Proof: The set $F = \{f^{(1)},f^{(2)},..,f^{(p)}\}$ is a strong complete set of alternating logic primitives if

$$T = \left\{f^{(1)},f^{(2)},...,f^{(p)},\bigcup_{m=1}^{p}\ \bigcup_{i=1}^{k_m} f_{0,i}^{(m)}\right\}$$

is complete in the conventional sense. And, by Theorem 2, T is complete if $T \not\subset M_i$ for $1 \leq i \leq 5$. Now, F contains a function $f^{(\ell)}$, $1 \leq \ell \leq p$, which is not 0-preserving by hypothesis. So, $T \not\subset M_1$. But if a self-dual function is not 0-preserving it is not 1-preserving. So, $T \not\subset M_2$. By Lemma 3, $T \not\subset M_3$ as every subfunction $f_{0,i}$ is not self-dual. Finally, by Lemma 4 and Lemma 5 $T \not\subset M_4$ and $T \not\subset M_5$ if $F \not\subset M_4$ and $F \not\subset M_5$. So, T is complete if $F \not\subset M_1, M_4, M_5$.

Q.E.D.

The restrictions on F for weak and strong completeness have thus been derived. However, it can be shown that weak and strong completeness are equivalent properties of a function set F of alternating primitives.

Theorem 4: Any weak complete set of alternating logic primitives is also a strong complete set of alternating logic primitives.

Proof: Assume a set $F = \{f^{(1)}, f^{(2)}, \ldots, f^{(p)}\}$ is a weak complete set of alternating logic primitives. Then $F \not\subset M_1, M_4$. But, if $F \not\subset M_1$, then there is a function $f^{(i)}$, $1 \leq i \leq p$, $f^{(i)} \varepsilon F$, having the property that $f^{(i)}$ is not 0-preserving. So,

$$f(0,0,\ldots,0) = 1$$

and by self-duality

$$f(1,1,\ldots,1) = 0.$$

But then there is a pair of n-tuples namely 000...0 and 111...1 with the following relationship

$$111\ldots1 > 000\ldots0$$

$$f(1,1,\ldots,1) = 0 < 1 = f(0,\ldots,0).$$

So, f is not monotonically increasing, and $F \not\subset M_5$. So, if $F \not\subset M_1, M_4$ then $F \not\subset M_1, M_4, M_5$.

Q.E.D.

Since weak and strong completeness are equivalent properties, in the following, complete by itself will be used to describe a set F which contains a function which is not 0-preserving and a function which is nonlinear.

SELF CHECKING CHARACTERISTICS OF COMPOSITIONS OF ALTERNATING LOGIC PRIMITIVES

The methods which have been described yield alternating circuits composed of alternating logic primitives representing any arbitrary Boolean function. These circuits, however, may not be self-checking. Consider an arbitrary alternating network N constructed as a composition of alternating logic primitives with n inputs $(x_1,\bar{x}_1),(x_2,\bar{x}_2),\ldots,(x_n,\bar{x}_n)$ and m outputs $(f_1,\bar{f}_1),(f_2,\bar{f}_2),\ldots,(f_m,\bar{f}_m)$. During fault-free operation inputs alternate and, as a result, the output of any logic primitive having only primary network inputs as inputs must alternate. But then, the output of any logic primitive having as inputs, primary network inputs or outputs from primitives having as input only primary network inputs, must alternate. Continuing, the output of every alternating logic primitive in a composition of alternating logic primitives must alternate if alternating inputs are applied. Faults, however, may occur in the network. Such faults are assumed to occur as stuck-at faults on input and output lines of alternating logic primitives. More formally, the fault set F is assumed to be the set of all single stuck-at-0 and stuck-at-1 faults on any alternating logic primitive input or output line. Assuming that a fault from the set F is present in the network N, the network is self-checking for that fault if no erroneous alternating output appears on f_i, $1 \le i \le m$, unaccompanied by a nonalternating output on some f_j, $1 \le j \le m$, $i \neq j$.

Theorem 5: Any network N consisting of a composition of alternating logic primitives is self-checking for all single stuck-at-0 and stuck-at-1 faults on any alternating logic primitive input or output line.

Proof: Assume a single fault f_a^d is present on an arbitrary logic primitive input or output line, a. Since the network N is a composition of alternating logic primitives, in fault-free operation either $(d,\bar{d})$ or $(\bar{d},d)$ is applied on a. If $(d,\bar{d})$ is applied then the fault is sensitized only during the application of $\bar{X} = \bar{x}_1\bar{x}_2\ldots\bar{x}_n$ to N. So, on any output f_i, $1 \le i \le m$, either (f_i,f_i) is produced if $\bar{X}$ also sensitizes a path from a to output f_i, or $(f_i, \bar{f}_i)$, the normal output, is produced if $\bar{X}$ does not sensitize a path from a to the output f_i. If $(\bar{d}, d)$ is applied then the fault is sensitized only during the application of $X = x_1 x_2 \ldots x_n$ to N. As a result, on any output f_i, $1 \leq i \leq m$ only $(f_i, \bar{f}_i)$ or $(\bar{f}_i, \bar{f}_i)$ is produced. In no case can an erroneous alternating output be produced and N is therefore self-checking for all single faults from the fault set F.

Q.E.D.

The significance of Theorem 5 lies in its generality. According to the theorem any alternating logic design structured as a composition of alternating primitives is self-checking for all single stuck-at faults regardless of structures. The advantages of this property will become obvious in the following section.

ALTERNATING DESIGN WITH THE MINORITY MODULE

Consider the set $F = \{f^{(1)}\}$, where $f^{(1)}$ is the three input minority gate. Now, $f^{(1)} \not\subset M_1$, $f^{(1)} \not\subset M_4$ so F is a complete set of alternating logic primitives. That is, the function $f^{(1)}$ is complete in itself being self-dual, but neither 0-preserving nor linear. If the (0,1) clock is applied on any input of the $\bar{M}$ module realizing $f^{(1)}$, then in an alternating network, the NAND of the other two inputs is realized.

$$f^{(1)} = \overline{(x_1 + x_2)(x_1 + x_3)(x_2 + x_3)}$$

$$f^{(1)}_{0,1} = \overline{x_2 x_3} \quad f^{(1)}_{0,2} = \overline{x_1 x_3} \quad f^{(1)}_{0,3} = \overline{x_1 x_2}$$

The $\bar{M}$ module configured with the (0,1) clock as one input will be referred to as the "alternating NAND."

One design procedure for obtaining alternating networks composed of minority gates proceeds as follows. The Boolean function to be realized in an alternating network is first realized as a conventional NAND network. Then, each NAND gate in this network is replaced by an alternating NAND to form an alternating network. Primary network inputs in the conventional design become alternating primary inputs in the alternating design, alternating in synchronism with the (0,1) clock, which is used as input to each alternating NAND. The result is an alternating circuit realizing f.

Example 1: Assume that a function f is to be realized in an alternating logic design using minority gates where f is given as

$$f = \overline{(\overline{x_2 \;\overline{x_2x_4}})(\overline{x_1 \;\overline{\overline{x_2x_4}\;\;\overline{x_1x_3}}})}.$$

Since f is already specified as a NAND decomposition, a conventional NAND network realizing f can be obtained from the functional specification directly. This network is shown in Figure 1 a). To obtain an alternating network f each NAND in the conventional network is first replaced with an $\bar{M}$ module having one input tied to the (0,1) clock which is assumed to be available. Primary inputs to the conventional network become synchronized alternating primary inputs to the alternating network. The resulting network is an alternating network realizing the function f and is shown in Figure 1 b).

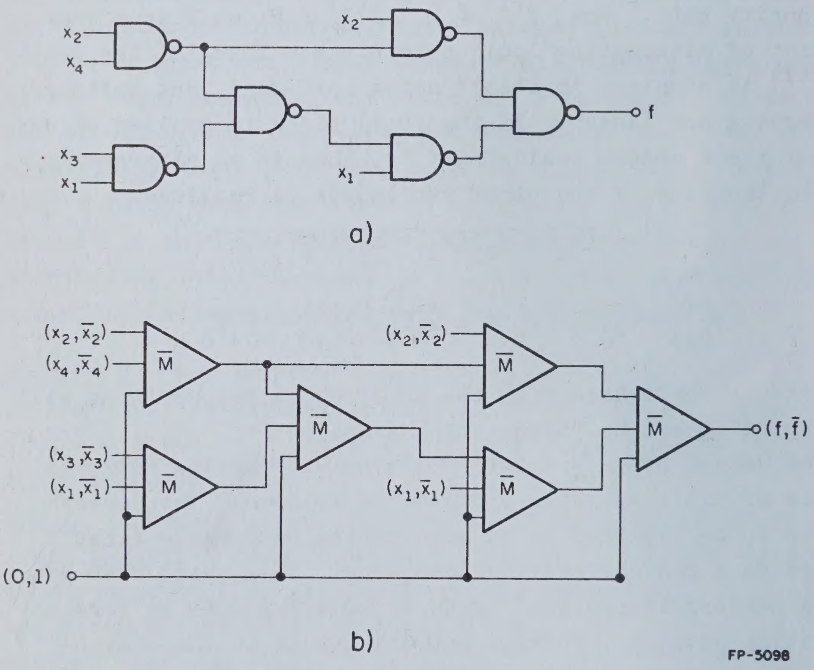


Figure 1 Minority Module Design

Further, the resulting alternating network is guaranteed to be self-checking even though the NAND network from which it was obtained has reconvergent fanout of unequal inversion parity, a structure prohibited in alternating design with conventional primitives [3,4].

A similar design procedure using the NOT gate and the 3-input majority gate can be described which uses conventional AND/OR/NOT designs as a starting point instead of NAND designs. This method is described in [4,5].

REFERENCES

1. Ibuki, K., Naemura, K., and Nozaki, A., "General Theory of Complete Sets of Logical Functions," Electronics and Communications in Japan (IEEE Translation), Vol. 46, No. 7, pp. 55-65, July, 1963.

2. Mukhopadhyay, A., "Complete Sets of Logic Primitives," Recent Developments in Switching Theory, Academic Press, pp. 1-26, 1971.

3. Reynolds, D., and Metze, G., "Fault Detection Capabilities of Alternating Logic," Proc. 6th Annual Symposium on Fault Tolerant Computing, pp. 157-162, June, 1976.

4. Reynolds, D., "The Design of Alternating Logic Systems with Fault Detection Capabilities," Coordinated Science Laboratory Report, University of Illinois, September, 1976.

5. Yamamoto, H., Watanabe, T., and Urano, Y., "Alternating Logic and Its Application to Fault Detection," Proc. 1970 IEEE International Computer Group Conference, Washington, D.C., pp. 220-228, June, 1970.

A FUNCTIONAL FORM APPROACH TO FAULT INTERRELATIONSHIPS IN TREE NETWORKS

VINOD K. AGARWAL and GERALD M. MASSON
Electrical Engineering Department
The Johns Hopkins University
Baltimore, Maryland 21218

ABSTRACT

To efficiently perform the fault analysis of tree networks, it is necessary that pertinent fault interrelationships be utilized. However, to determine them can entail an analysis which is quite complex and reduces the overall advantage of utilizing the gained insights in a fault analysis process. In this paper we suggest an approach to the establishing of the existence of certain fault interrelationships in tree networks which is based only on the form of the output expression. Rules are given for generating a form complex corresponding to that expression. Theorems are stated regarding the interpretation of those form complexes. With these results, some well-known multiple fault covering observations are seen to be easily extendable.

I. Introduction

The fault analysis of combinational switching networks can be completely based on the algebraic manipulation of Boolean expressions relevant to that network. Hence, it has been found that techniques and concepts which were originally developed for the logical design of such networks and which are also algebraically manipulative in nature also prove useful in fault analysis. Indeed, significant examples of such joint utilizations exist, the most notable, perhaps, being the functional equivalence of Boolean expressions [1]; the corresponding functional equivalence of faults is commonly employed by fault analysis researchers and has led to many important results. We, therefore, have been led to the consideration from a fault analysis point of view of other equivalence relations which can be defined on the set B(n) of all 2^{2^n} Boolean expressions of n variables. In particular in this paper, we develop a form equivalence concept for Boolean expressions and define a corresponding form equivalence among faults in combinational tree networks. Since the form of a Boolean expression does not contain the same information as an explicit functional expression , we must expect that we will not be able to obtain detailed tests based on it. However, it will be seen that in certain situations the form is sufficient to indicate certain fault interrelationships and the existence of certain fault phenomena. This information can then be used with explicit functional expressions to more efficiently perform fault analysis.

II. Preliminaries

We will denote, in general, an n-input, single-output combinational switching network as N. This network will consist only of interconnected AND, OR, NAND, NOR, and NOT gates, and each line of N will be labelled with a distinct integer. (For example, see Figure 1). X will denote an n-tuple binary vector, $(X_1,\ldots, X_n)$, $X_i \in \{0,1\}$, $i=1,\ldots, n$, which corresponds to the primary inputs of N. (The convention $X_i^0 = \overline{X}_i$ and $x_i^1 = X_i$ will sometimes be used in the following). Y_j will denote a Boolean expression describing line j of N in terms of the primary inputs. We will refer to Y_j as the internal line variable for line j.

The capital letters I, J, K,... will be used to denote set of integers corresponding to subsets of the labelled lines of N. We use these as subscripts to succinctly describe the set of Boolean expressions corresponding to the internal line variables of, say, I as $Y_I = \{Y_1,\ldots, Y_p\}$, where, say, $I = \{1,\ldots, p\}$. When we are considering only the form of an expression, and are not specifically concerned with whether the internal line variable of, say, line $i \in I$ is explicitly present in either a complemented or uncomplemented form in the expression, we will denote the line variable as y_i. This generalizes directly to y_I for the set of internal line variables corresponding to I. Finally, we will denote the primary output Boolean expression of N as $Z[X;Y_I]$ when it is written in terms of X and Y_I.

A fault condition corresponding to line i of N stuck at α_i, $\alpha_i \in \{0,1\}$, will be denoted as $f_i^{\alpha_i}$. Similarly, $f_I^{\alpha_I}$ will denote the set of stuck-type faults $\{f_1^{\alpha_1},\ldots, f_p^{\alpha_p}\}$, where, say, $I = \{1,\ldots, p\}$. The primary output expression of the network with $f_I^{\alpha_I}$ present is then written as $Z[X;\alpha_I]$. Clearly, this is obtained from $Z[X;Y_I]$ by substituting $\alpha_i \in \alpha_I$ for $Y_i \in Y_I$ for all $i \in I$.

Finally, the subset of all binary input vectors which represents the standard complete test set for $f_I^{\alpha_I}$ will be denoted as $T(f_I^{\alpha_I})$.

There are many techniques by which $T(f_I^{\alpha_I})$ can be generated, but an algebraically manipulative method which is most relevant to the above discussion is the Boolean difference technique [2]-[4] wherein

$$T(f_I^{\alpha_I}) = \{X \mid Z[X;Y_I] \oplus Z[X;\alpha_I]=1\} \qquad (2\text{-}1)$$

Many interesting results have been based on the Boolean difference concept and its various expansions. It is interesting to note the prominance of $Z[X;Y_I]$ in the generation of the test set for $f_I^{\alpha_I}$. This suggests that certain properties of $Z[X;Y_I]$ might have useful ramifications relative to the test set $T(f_I^{\alpha_I})$. Indeed, the principle point of the paper is to describe such a property and indicate some of its ramification. More particularly, the property which we will consider will be functional form and the ramification will involve the existence of an interesting type of multiple fault, called the singly noncovered multiple fault, henceforth denoted as an SNM fault. Consider $I=\{1,\ldots, p\}$, $T(f_i^{\alpha_i})$, $i \varepsilon I$, and $T(f_I^{\alpha_I})$. It is usually the case that $T(f_i^{\alpha_i})$ (that is, a single fault test set) is computationally simpler to determine than $T(f_I^{\alpha_I})$ (that is, the multiple fault test set). Moreover, there are certainly far fewer single fault possibilities in a combinational switching network than single and multiple fault possibilities. Hence, if a test set which detected all $f_i^{\alpha_i}$, $i \varepsilon I$, were generated, it would be significant to know which (if any) multiple faults on the lines of I might not be detected by that single fault test set. If T_{cs} denotes a complete single fault test for the p single faults $f_i^{\alpha_i}$, $i \varepsilon I = \{1,\ldots,p\}$, then $T(f_i^{\alpha_i}) \cap T_{cs} \neq \phi$ for all $i \varepsilon I$. But, it is possible that $T(f_I^{\alpha_I}) \cap T_{cs} = \phi$. In other words, even though, T_{cs} covers each single fault $f_i^{\alpha_i}$ associated with the lines of I, it may not cover the multiple fault where each of those faults on I occur simultaneously. $f_I^{\alpha_I}$ is, therefore, called a SNM fault.

The non existence of SNM faults would certainly simplify multiple fault analysis. Indeed, results involving the existence or nonexistence of SNM faults for certain type of combinational switching networks are known [5]-[6]. In the following we will duplicate and extend some of these results using a functional form approach. This will be seen to add some new insights into the interaction of faults in combinational switching networks.

III. Forms of Boolean Expressions

In the tree network N, for any single line i, we have the following:

Lemma 3.1

$$Z[X;Y_i] = A_i y_i + B_i$$

where A_i and B_i are expressions consisting only of primary input variables, and 0 and 1, and $y_i \varepsilon \{Y_i, \overline{Y}_i\}$.

Proof: Clearly, the more general form, $Z[X;Y_i] = A_i y_i + A_i' \overline{y}_i + B_i$ could never result from a tree network since there are no fanout paths.

Q.E.D.

Hence, we will, in the following, be regularly dealing with expressions of this form. It is, therefore, expedient to state the following:

Definition 3.1: A Boolean expression $Z[X;Y_i]=A_i y_i+B_i$ as detailed in Lemma 3.1 is said to describe a form F, and this is denoted $Z[X;Y_i] \underset{d}{=} F$.

It should be noted that this association of F with $Z[X;Y_i]$ is independent of the particular A_i and B_i involved (and, indeed, includes the special case of $A_i=1$ and $B_i=0$).

Thus, we now have a basic form which corresponds to the output function of N in terms of X and some Y_i. When we consider some set of lines of N, say, I, $|I|>1$, the resulting form of $Z[X;Y_I]$ will be of interest. Such a form will be denoted by a combination of F's, interconnected by "+" and "•" operators. We will refer to this form as the F-complex for $Z[X;Y_I]$. The following will lead to the generation of F-complexes.

Lemma 3.2: If $Z[X;Y_i] \underset{d}{=} F$, then

(i) $\overline{Z[X;Y_i]} \underset{d}{=} F$

(ii) $Z[X;Y_i] \cdot C_i \underset{d}{=} F$

(iii) $Z[X;Y_i] + C_i \underset{d}{\equiv} F$

where, for parts (ii) and (iii), C_i is an expression in the primary input variables.

Proof: The lemma follows directly from DeMorgan's laws and/or composition of the expressions A_i, B_i, and C_i (and their complements).Q.E.D.

Consider now two subtrees of N, say N_1 and N_2, and two lines i_1 and i_2 of N_1 and N_2, respectively. By Lemma 3.1, we know that $Z_1[X;Y_{i_1}] \underset{d}{=} F$ and $Z_2[X;Y_{i_2}] \underset{d}{=} F$. To determine the form described by $Z[X;Y_I]$, where $I=\{i_1,i_2\}$, we need composition rules. Note that initially we are only considering here two Boolean expressions, both of which describe the form F. It should also be noted that such rules must be consistent with Definition 3.1 and utilize Lemma 3.2.

Example NAND Rule:

$$Z[X;Y_I] = \overline{Z_1[X;Y_{i_1}] \cdot Z_2[X;Y_{i_2}]}$$

$$\longleftrightarrow \quad Z[X;Y_I] \underset{d}{=} \overline{F \cdot F}$$

Similar rules can be easily obtained for the OR, AND, and NOR operations on $Z_1[X;Y_{i_1}]$ and $Z_2[X;Y_{i_2}]$, but, for conciseness we will not state them here.

Example Complementation Rule: Let H be the F-complex F+F $Z[X;Y_I \underset{d}{\equiv} H \longleftrightarrow \overline{Z[X;Y_I]} \underset{d}{=} \overline{H}$, where $\overline{H}$ is the F-complex F•F.

Next, we give a lemma and an example rule for generating an F-complex based on $Z_1[X;Y_{I_1}] \underset{d}{=} H_1$, $|I_1| \geq 1$, and $Z_2[X;Y_{I_2}] \underset{d}{=} H_2$, $|I_2| \geq 1$, where $I_1 \cap I_2 = \emptyset$, and $I_1 \cup I_2 = I$.

Lemma 3.3: If $Z[X;Y_{I_j}] \underset{d}{=} H_j$, then (i) $Z[X;Y_{I_j}] \cdot C_{I_j} \underset{d}{\equiv} H_j$ and (ii) $Z[X;Y_{I_j}] + C_{I_j} \underset{d}{=} H_j$, where C_{I_j} is any expression in the primary input variables.

Proof: The lemma follows from noting that for any complex of F's described by $Z[X;Y_{I_j}]$, the C_{I_j} will be absorbed by variables in the complex which also only consist of primary input variables. Q.E.D.

Example General NAND Rule:

$$Z[X;Y_I] = \overline{Z_1[X;Y_{I_1}] \cdot Z_2[X;Y_{I_2}]}$$

$$\longleftrightarrow Z[X;Y_I] \underset{d}{=} \overline{H_1 \cdot H_2}$$

Given two F-complexes, say, H_1 and H_2. If these complexes can be manipulated so that the resulting F-complexes are identical, then they are said to be equivalent F-complexes, and we denote this as $H_1 \underset{f}{=} H_2$. The manipulations allowed are given in the following rule:

Manipulation Rule: The operations "+" and "•" in F-complexes are associative and commutative, but are not distributive. Thus, for example, $F \cdot (F+F) \underset{f}{=} (F+F) \cdot F$, and $F+(F+F) \underset{f}{=} F+F+F$, but $F \cdot (F+F) \underset{f}{\neq} F \cdot F + F \cdot F$.

All these lemmas, rules, and definitions can be recursively used to obtain the F-complex described by $Z[X;Y_I]$. The following two lemmas follow:

Lemma 3.4: In a tree network, $Z[X;Y_I]$, $|I| = 2$, always describes either the form $F + F$ or the form $F \cdot F$.

Lemma 3.5: In a tree network, $Z[X;Y_I]$, $|I| = 3$, always describes one of the four forms: $F+F+F$, $F+(F \cdot F)$, $F \cdot (F+F)$, and $F \cdot F \cdot F$.

IV. An Example of Form Utilization

In Section II, we noted that it is often useful in multiple fault analysis to determine the non-existence of SNM faults in a given network. In this section, we will state a fundamental theorem which will relate the non-existence of SNM faults on certain lines to the existence of certain F-complexes described by the output expression in terms of those lines. Based on these results, a particularly interesting network will be described. To do this, we will first adapt the form concept for Boolean expressions to a form concept for faults.

Definition 4.1: A fault $f_I^{\alpha_I}$, $|I| = k \geq 1$, is said to describe the F-complex, H, when $Z[X;Y_I] \underset{d}{=} H$.

According to this definition all 2^k different faults on a particular set of I, $|I| = k$, lines describe the same F-complex Thus we will say that all 2^k such faults are form equivalent. Furthermore if two output expressions, say, $Z[X;Y_I]$ and $Z[X;Y_J]$, $I \cap J = \emptyset$, $|I| = |J| = k$, are form equivalent, then all 2×2^k faults on these lines are form equivalent. To state the fundamental theorem relating forms of faults and the SNM fault phenomena, we consider the following.

Let $f_I^{\alpha I}$ be a fault which describes the F-complex, say, H. Let T_c be a test set which can detect every subfault $f_{I'}^{\alpha I'}$, $I' \subseteq I$. Now, suppose that an additional single fault $f_j^{\alpha j}$, $j \notin I$, occurs in the network such that the composite fault is $f_J^{\alpha J}$, $J = I \cup \{j\}$. Let X_1 be a test input vector which can detect the single fault $f_j^{\alpha j}$. We can now state the following fundamental theorem:

Theorem 4.1: The composite fault $f_J^{\alpha J}$ and all of the subfaults $f_{J'}^{\alpha J'}$, $J' \subset J$, are detectable by the test set $T_c \cup \{X_1\}$, if the F-complex described by the composite fault is either H+F or H•F.

Proof: The details [15] of this proof require notational developments and lemmas describing a consistency property of the basic form F which while reasonably straightforward are descriptively lengthy, and, hence, will not be given here.

It should be pointed out that a similar theorem could be developed for other types of fault interrelationships such as minimal multiple redundancy [10],[13],or minimal complete masking [14]. We will now illustrate a recursive use of Theorem 4.1 and SNM fault phenomena.

For the conditions of the theorem, let $|I| = 1$, and let T_c be a complete single fault test set for a given N. Then $H \underset{f}{=} F$. Therefore, corresponding to any double fault, $f_J^{\alpha J}$, $|J|=2$, we have by Definition 4.1 and Lemma 3.4, that the F-complex described by the double fault is F+F or F•F. This satisfies the theorem, and thus we have proven that no multiple fault of size two in a tree network can be a SNM fault. Repeating the above, if we let $|I| = 2$, then $H \underset{f}{=}$ F+F or F•F. Therefore, corresponding to any 3-fault, $f_J^{\alpha J}$, $|J|=3$, we have, by Definition 4.1 and Lemma 3.5, that the F-complex described by the 3-fault is either H+F or H•F. Thus, no 3-fault in a tree network can be a SNM fault. These results about tree networks are, of course, known [5], [6]. However, the theory and arguments used here are more general. Continuing, we can certainly determine which (if any) K-faults, $K \geq 4$, might be SNM faults. Note, for example, that in the case of 4-faults, the only possibilities for SNM faults are those which describe either the F-complex (F+F)•(F+F) or (F•F)+(F•F); Gault, et. al [5] showed such a SNM fault whose F-complex was (F+F)•(F+F). Also, by similarly using Theorem 4.1 in tree networks such as two level and cascaded two level networks [6] (also called restricted fanout-free networks) it can be shown that no fault $f_I^{\alpha I}$, $I \geq 2$ can be a SNM fault.

Another interesting utilization of Theorem 4.1 will now be illustrated regarding a class of networks which we will call $(a_0,a_1,a_2,\ldots,a_k)$-networks, where $a_0=1<a_1<a_2<\ldots<a_k$, $k \geq 0$, are some k integers. Such networks will be

defined to have the property that a test set designed to detect every a_i-fault, $i=0,1,\ldots,k$, in the network will always detect every other multiple fault as well. For example, every two level and cascaded two level network is a (1)-network. As another example, consider again the network of Figure 1. It has been shown [5] that except for the particular 4-fault, $\{f_1^1, f_3^0, f_5^1, f_7^0\}$, the complete single fault test set indicated in the figure, will detect every multiple fault of size less than or equal to 4. (Indeed, this fault describes the F-complex (F+F)•(F+F)). Moreover, it is easily argued that every multiple fault of size greater than 4 in this network is functionally equivalent to some K-fault, $K \leq 4$, it is clear that if we augment the complete single fault test set T_{cs} one additional input vector generated specifically to test the fault $\{f_1^1, f_3^0, f_5^1, f_7^0\}$, then the resulting test set will definitely detect every single or multiple fault in this network. Hence, the network of Figure 1 is a (1,4) network. We will illustrate here how the network of Figure 1 can be modified to obtain a $(1,a_1,a_2,\ldots,a_k)$ network, where $a_i=2(i+1), i=1,2,\ldots,k$ for arbitrary k.

Thus, consider the network, N, of Figure 2[1]. Let T_{cs} be a test set which detects every single fault in N. By Theorem 4.1, T_{cs} will also detect every 2-fault and every 3-fault. However, there are 4-faults in this network which describe the F-complex $H \overset{=}{f} (F+F) \cdot (F+F)$, and, as discussed earlier, such 4-faults may not be detected by T_{cs}. Suppose, then, we add to T_{cs} test vectors which will test for all 4-faults. The resulting test set, denoted as T_{cs}', will then detect all single and multiple faults up to 4-faults. Note now that all 5-faults describe F-complex of the type H+F or H•F. By Theorem 4.1 they are therefore definitely covered by T_{cs}'. However, this is not the case for all 6-faults. For example, $\{f_1^0, f_3^1, f_5^0, f_7^1, f_9^0, f_{11}^1\}$ describes the F-complex (F+F)•(F+F)•(F+F). By Theorem 4.1 such a 6-fault may not be covered by T_{cs}'. In a similar manner to the above, we then augment T_{cs}' to also cover all 6-faults. Thus, the network of Figure 2 is a (1,4,6)-network. Extensions to (1,4,6,8,...)-networks should be clear.

---

[1] This particular (1,4,6)-network is due to Mr. Shinji Nakamura, Electrical Engineering Department, The Johns Hopkins University.

V. Discussion

In this paper we have considered the determining of fault interrelationships in tree networks by means of form descriptions for the output expression of the network. It was shown that if certain rules are followed in generating the F-complexes for an output expression, we can, indeed, state conclusions regarding fault phenomena. We chose to demonstrate this by using singly noncovered multiple faults. It is emphasized that conclusions regarding other fault phenomena could have been stated as well [15]. In addition, the consistency property mentioned in the proof of Theorem 4.1 can be defined for non-tree networks also. Therefore, it is possible to entend somewhat [15] the results described here to non-tree networks.

ACKNOWLEDGEMENT

The authors would like to acknowledge the Office of Naval Research Contract N000014-75-C-119 for the support of this research.

REFERENCES

[1] R.E. Prather, "Introduction to Switching Theory: A Mathematical Approach". Published by Allyn and Bacon, Inc., Boston, Mass.

[2] S.S. Yau and Y.S. Tang, "An efficient algorithm for generating complete test sets for combinational logic circuits", IEEE Trans. Comput., Vol. C-20, pp. 1245-1251, November, 1971.

[3] C.T. Ku and G.M. Masson, "The Boolean difference and multiple fault analysis", IEEE Trans. Comput., Vol. C-24, pp. 62-71, January, 1975.

[4] V.K. Agarwal and G.M. Masson, "Resolution oriented fault interrelationship in combinational logic network", to appear in IEEE Trans. Comput.

[5] J.W. Gault, J.P. Robinson, and S.M. Reddy, "Multiple fault detection in combinational networks", IEEE Trans. Comput., Vol. C-21, pp.31-36, January, 1972.

[6] D.R. Schertz and G. Metze, "On the design of multiple fault diagnosable networks", IEEE Trans. Comput., Vol C-20, pp. 1361-1364, November, 1971.

[7] E.J.McCluskey and F.W. Clegg, "Fault equivalence in combinational logic networks", IEEE Trans. Comput., Vol. C-20, pp. 1286-1293, November, 1971.

[8] A.D. Friedman, "Fault detection in redundant circuits", IEEE Trans. Electron. Comput., Vol. EC-16, pp. 99-100 January, 1967.

[9] R. Dandapani and S.M. Reddy, "On the design of logic networks with redundancy and testability considerations", IEEE Trans. Comput., Vol. C-23, pp. 1139-1149, November, 1974.

[10] H.P.S. Lee and E.S. Davidson, "Redundancy testing in combinational networks", IEEE Trans. Comput., Vol. C-23, pp. 1029-1047, October, 1974.

[11] J.E. Smith and G. Metze, "On the existence of combinational networks with arbitrary multiple redundancies", Proc. Thirteenth Annual Allerton Conf. on Circuit and System Theory, Monticello, Ill., October, 1975, pp. 692-703.

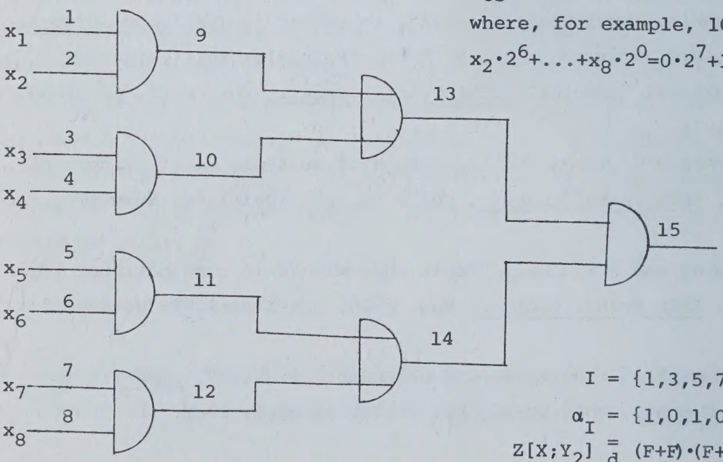
[12] G. Metze and C.W.Y. Cha, "Multiple fault diagnosis in combinational networks", Proc. Twelfth Annual Allerton Conf. on Circuit and System Theory, Monticello, Ill., October 1974, pp. 244-253.

[13] V.K. Agarwal and G.M. Masson, "Some redundancy and complete masking observations", short paper, in this proceedings.

[14] V.K. Agarwal and G.M. Masson, "On redundancy phenomena and complete masking in combinational networks", in review IEEE Trans. Comput., also The Johns Hopkins University, Baltimore, Md., JHU-EE Report No. 75-27, May, 1976.

[15] V.K. Agarwal and G.M. Masson, "A functional form approach to fault interrelationships in switching networks", Department of Electrical Engineering, The Johns Hopkins University, Baltimore, Md., JHU-EE Report No. 76-3.

A Complete Single Fault Test Set T_{cs}= {107,126,182,215,155,187}, where, for example, $107=x_1 \cdot 2^7 + x_2 \cdot 2^6+\dots+x_8 \cdot 2^0=0\cdot 2^7+1\cdot 2^6+\dots+1\cdot 2^0$.

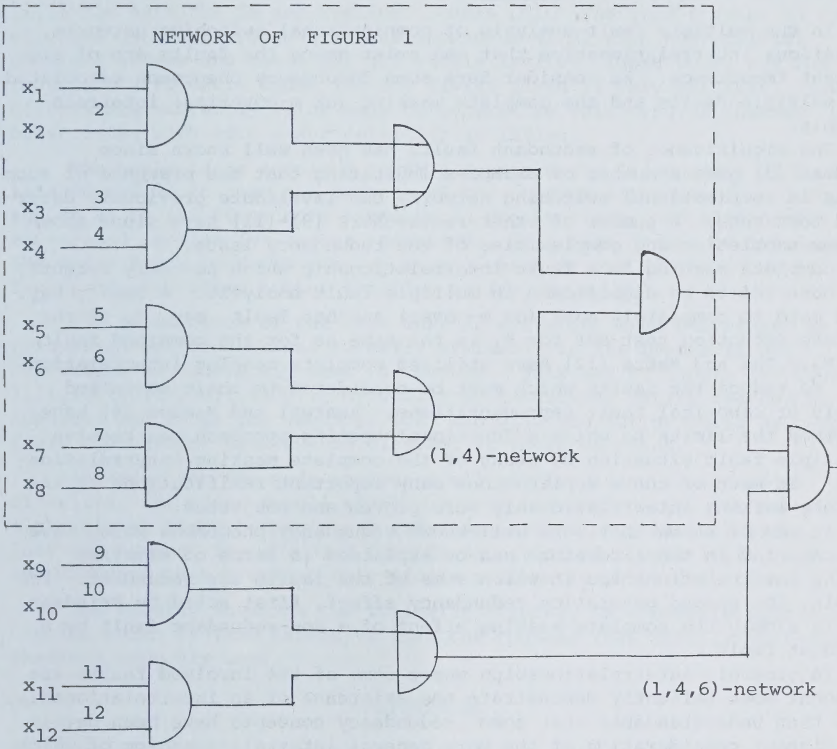


$I = \{1,3,5,7\}$

$\alpha_I = \{1,0,1,0\}$

$Z[X;Y_2] \underset{d}{=} (F+F)\cdot(F+F)$

FIGURE 1


NETWORK OF FIGURE 1
x_1
1
x_2
2
x_3
3
x_4
4
x_5
5
x_6
6
x_7
7
x_8
8
(1,4)-network
x_9
9
x_{10}
10
x_{11}
11
x_{12}
12
(1,4,6)-network

$$I = \{1,3,5,7,9,11\}$$

$$\alpha_I = \{0,1,0,1,0,1\}$$

$$Z[X;Y_I] = (x_2y_1 + x_4y_3)(x_6y_5 + x_8y_7)(x_{10}y_9 + x_{12}y_{11})$$

$$\underset{d}{=} (F+F)\cdot(F+F)\cdot(F+F)$$

FIGURE 2

SOME REDUNDANCY AND COMPLETE MASKING OBSERVATIONS [1,2]

VINOD K. AGARWAL and GERALD M. MASSON
Electrical Engineering Department
The Johns Hopkins University
Baltimore, Maryland 21218

SUMMARY

In the multiple fault analysis of combinational switching networks, the various interrelationships that can exist among the faults are of significant importance. We consider here some redundancy phenomena associated with multiple faults and the complete masking (or m-covering) interrelationship.

The significance of redundant faults has been well known since Friedman [8] gave a number of examples indicating that the presence of such faults in combinational switching networks can invalidate previously determined test sets. A number of other researchers [9]-[11] have since shown further subtleties and complexities of the redundancy issue.

Complete masking is a fault interrelationship which has only recently been observed to be significant in multiple fault analysis. A fault, say, F_1 is said to completely mask (or m-cover) another fault, say, F_2 if the complete detection test set for F_1 is the same as for the combined fault $F_1 \cup F_2$. Cha and Metze [12] have utilized complete masking interrelationship to reduce the faults which must be considered to their so-called kernels or canonical fault representations. Agarwal and Masson [4] have described the limits to which a functional testing approach can resolve a multiple fault situation by means of the complete masking interrelationship. In each of these applications many important ramifications of the complete masking interrelationship were proven and conjected.

It can be shown that some well-known redundancy phenomena which have been reported in the literature can be explained in terms of complete masking interrelationships in which some of the faults are redundant. For example, the second generation redundancy effect, first noted by Friedman [8], is simply the complete masking effect of a non-redundant fault by a redundant fault.

In general, interrelationships where some of the involved faults are redundant most saliently demonstrate the existence of an interrelationship. It is then understandable that some redundancy concepts have been developed without consideration of the more general interrelationships of which they are indicative. To exemplify this, we have the following theorem:

Theorem: If a fault, say, F_1 completely masks another fault, say, F_2 in a network, then in the faulty network where F_1 is present, fault F_2 is a redundant fault. Conversely, for every redundant fault, say, F_2 in a network, there exists another network and a fault, say, F_1 in that new network such that F_1 completely masks F_2.

This theorem is suggestive of the existence of a general theory of fault interrelationships relating the functional and structural characteristics of a network.

---

1. This research was supported by the Office of Naval Research Conract N00014-75-C-119.
2. All references correspond to those of "A functional form approach to fault interrelationships in tree networks", by V.K. Agarwal and G.M. Masson, this proceedings.

A LOW COST TOTALLY SELF-CHECKING CHECKER FOR RESIDUE 3 AND 3N CODES*

JEAN DUSSAULT and GERNOT METZE
University of Illinois

This short paper introduces a low-cost totally self-checking checker [1,2] for both the 3N and residue 3 codes [3]. The tree checker is constructed of blocks which are less expensive (2/3) than conventional modulo 3 adders and check for all single faults. These codes are the least redundant arithmetic codes and the design given is very specific. However efforts are currently being made to generalize this type of checker to other codes with more error detecting abilities.

The basic ideas involved in this design are: 1) encoding of 0 mod 3 into 00 and 11 so that the outputs of the checker alternate more frequently and so as to introduce more symmetry in the function; 2) elimination of end-around carries for faster operation; 3) a high degree of modularity: the basic block performs $F(x_1,x_2,x_3,x_4)$ which is one of the 2 output functions; the other output function is $F(x_2,x_1,x_4,x_3)$.

The final block of the tree must be different from the others as the restrained set of vertices normally available to its inputs is not sufficient to completely test the module. This set of inputs is {0000, 1111,1001,0110,0011,1100}. A totally self-checking parity checker is adequate and since the vertices 1001 and 0110 are not produced by a modeled fault the checker need not be disjoint in those vertices.

In order for the code to detect all single faults in the adder circuitry, the adder itself should produce independent sum and carry. The minimal such circuit requires 9 gates against 6 for a minimal adder. This last requirement is sufficient for the detection of single logic faults for any encoding that provides only single arithmetic error detection.

The result, shown in the following figures, is a totally self-checking checker which is less expensive than the regular (not totally self-checking) checkers commonly used.

REFERENCES

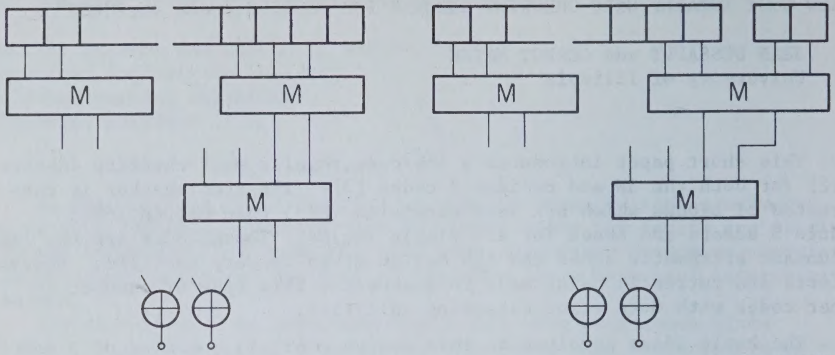
[1] Anderson, D. A., "Design of Self-Checking Digital Networks Using Coding Techniques," Coordinated Science Laboratory Report R-527, Univ. of Illinois, Urbana, Sept. 1971.

[2] Smith, J. E., "The Design of Self-Checking Combinational Circuits," Coordinated Science Laboratory Report R-737, Univ. of Illinois, Aug. 1976.

[3] Avizienis, A., "Arithmetic Codes: Cost and Effectiveness Studies for Application in Digital System Design," IEEE Trans. Computers, Vol. C-20, Nov. 1971, pp. 1322-1331.

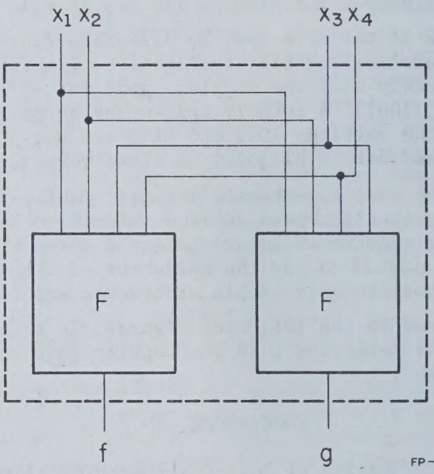
---

*This work was supported by the Department of Education of the Province of Quebec and in part by the Joint Services Electronics Program (U. S. Army, U. S. Navy, and U. S. Air Force) under Contract DAAB-07-72-C-0259.



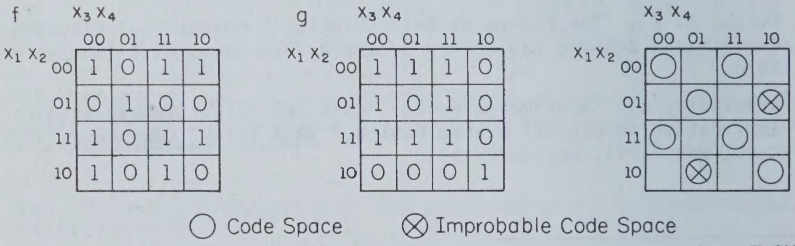
M
M
M
M
M
M
FP-5219

Figure 1. General Structure of the Checker for 3N and Residue 3 Codes.



x_1 x_2
x_3 x_4
F
F
f
g
FP-5220

Figure 2. M-Module.



Code Space
Improbable Code Space
FP-5221

f

x_1x_2 \ x_3x_4	00	01	11	10
00	1	0	1	1
01	0	1	0	0
11	1	0	1	1
10	1	0	1	0

g

x_1x_2 \ x_3x_4	00	01	11	10
00	1	1	1	0
01	1	0	1	0
11	1	1	1	0
10	0	0	0	1

x_1x_2 \ x_3x_4	00	01	11	10
00	○		○	
01		○		⊗
11	○		○	
10		⊗		○

○ Code Space ⊗ Improbable Code Space

Figure 3. Truth Table of $f(F(x_1,x_2,x_3,x_4)),g(F(x_2,x_1,x_4,x_3))$ and of the Final EX-ORs.

DESIGN OF ASYNCHRONOUS CIRCUITS — SOME PROBLEMS*

JON G. KUHL and S. M. REDDY
Division of Information Engineering
The University of Iowa
Iowa City, Iowa 52242

ABSTRACT

A review of some available techniques for the design of asynchronous circuits for use in fault-tolerant environments is done and several potential problems in their application are pointed out.

I. INTRODUCTION

The design of asynchronous sequential circuits for use in fault-tolerant environments has been a subject of increasing attention recently. Several design techniques have appeared in the literature [1-5]. However, careful investigation shows many of these techniques to suffer from potentially severe problems. The purpose of this paper is to discuss some of these shortcomings.

Assumptions

Throughout this paper we assume stuck-at-logical constant fault model, with only single faults being considered. Faults are allowed at gate inputs or outputs, whereas the primary inputs are assumed to be fault-free. This fault model coincides with that assumed in the previous works discussed in this paper [1-3].

The asynchronous machines considered are assumed to be represented by normal flow tables and the asynchronous circuits are to operate in fundamental mode. In discussing the material in this paper we assume that only single input variables change at any time (all previous designs discussed herein implicitly assume this [1-3]).

Throughout this paper we assume that the reader is familiar with the terminology, definitions and design techniques given in the earlier works to be discussed [1-3].

---

*The research reported has been supported by NSF Grant ENG72-04142 A02

II. ASYNCHRONOUS CIRCUITS DESIGNED FOR FAULT DETECTION

Sawin and Maki have given a technique to design asynchronous circuits for fault detection [1]. We will show two potential problem areas arising in the application of their results.

(a) Test sequences to detect faults. Sawin and Maki have claimed that for their design: "entering any part of the transition path exercises the logic for the complete set of states which compose that particular transition path" [1]. A transition path is defined to be the set of states a machine may enter in the course of a transition from a present state to a next state. This would certainly imply, then, that entering every stable total state (a total state is the composite of the input state and the state of the asynchronous machine) in the flow table should be sufficient to test for all single faults. This is not true, however, as illustrated by the example in Figure 1(a).

The state code and the design equations in Figure 1(a) are in accordance with Sawin and Maki's technique [1]. Suppose the $\bar{x}_2$ input to y_1I_1 becomes s-a-1 (assuming that y_1I_1 is realized only once, by an AND gate, in the circuit). The next state equations have now become:

$$Y_1 = y_1(I_1 + I_2) + I_3$$

$$Y_3 = y_1(I_1 + I_2) + y_2I_1 + y_3I_2 + I_3$$

$$Y_5 = y_1(I_1 + I_2) + y_4I_2 + I_3 + y_5I_4$$

Y_2, Y_4, Y_6 are unchanged.

To detect this fault it is necessary to enter I_2 with $y_1 = 1$ in either the initial or final state of the transition. The only transition satisfying this condition is the one originating in A under I_1. But it is possible to cycle through all stable total states without entering this transition, as shown in Figure 1(b).

Thus Sawin and Maki's claim is clearly incorrect and, in fact, for a worst case situation, it might be necessary to enter all, or almost all transitions in the table to assure detection of all single faults.

To point out the potential seriousness of this problem, suppose that subsequent to the occurrence of the fault discussed above, but prior to its detection, a second fault causes the y_3I_2 AND gate to become s-a-0. Now if the circuit is stable in A under I_1 and the input changes to I_2, the stable state reached is A = 101010, instead of the proper state C = 011001.

Thus if Sawin and Maki's design is to have a high probability of detecting single faults for all flow table configurations the mean-time-between-failure (MTBF) must be assumed to be much greater than the time required to ensure the entering of all transitions during normal operation. For very large flow tables such an assumption might prove unacceptable.

Note: In fairness, it should be pointed out that Sawin and Maki made no claim that their design would be valid in the presence of multiple faults. None-the-less the example is important in that it shows the necessity of adopting a potentially unacceptable MTBF assumption in order to make their design technique valid.

(b) Disjoint logic and critical races. Consider the flow table portion shown in Figure 2. y_{j_i} is the ith controlling variable for input I_j. Suppose that separate AND gates realizing $y_{P_1} I_P$ are used to feed to the OR gates realizing Y_{P_1} and Y_{R_2}. Sawin and Maki do not prohibit the use of disjoint logic and, in fact, specifically discuss how to connect the feedback paths in this situation on page 247 of reference 1. But now assume that the $y_{P_1} I_P$ gate feeding to Y_{R_2} suffers a stuck-at-1 fault on an input lead such that it realizes y_{P_1} $(I_P + I_R)$. Consider the sequence of events occurring if the machine is in A under I_Q and input changes to I_R. The faulty AND gate will excite and hence Y_{R_2} becomes 1. But at the same time there is no longer any term holding Y_{P_1} to 1. If y_{P_1} becomes 0 sufficently quickly, then the faulty term might go to 0 and subsequently Y_{R_2} will go to zero also. The circuit will thus assume its proper stable state (under the control of y_{R_1}).

However, suppose Y_{R_2} remains 1 sufficiently long for y_{R_2} to become 1. Then the $y_{R_2} I_R$ term will excite and hold Y_{R_2} to 1. Now both y_{R_1} and y_{R_2} are 1 and this is clearly a detectable condition.

Thus, the detectability of the fault hinges on the outcome of a critical race between the 1/0 transition of Y_{P_1} and the 0/1 transition of Y_{R_2}. We call this occurrence a "pseudo-race" since it is due to the disjoint nature of the logic rather than any inherent property of the next-state equations.

This problem is serious since it implies that the detectability of certain faults depends on the nature of stray delays in the circuit, over which the logic designer might not have any controls. To illustrate a potential disasterous consequence of pseudo-races, let us return to Figure 2 under the fault conditions described above. Assume that stray delays are such that the circuit operates in a way making the fault undetectable. Suppose that at some later time the $y_{R_1}I_R$ term feeding to Y_{R_1} becomes s-a-0 and further suppose that the nature of the stray delays have changed such that the circuit now operates as to make the first fault detectable (since stray delays might depend on the gate loads at a given time this situation could occur). Now if the circuit is in A under I_Q and input changes to I_R, Y_{R_2} becomes 1 due to the first fault and Y_{R_1} goes to 0 due to the second fault. The final state reached is then G which in an incorrect but proper stable state.

The example illustrates that the use of shared logic may, in fact, be essential to Sawin and Maki's design in order to prevent critical race situations. It is thus implied that the proposed design [1] may not be applicable in situations where the use of disjoint logic is required (e.g. certain logic array implementations, or instances where device technology does not allow sufficient fan-out capability).

III. FAULT-TOLERANT ASYNCHRONOUS MACHINES

Sawin and Maki have proposed a fault tolerant asynchronous circuit employing two copies of the circuit using the design described in the previous section [2]. The two circuits operate simultaneously and the parity of the partitioning variables for the currently valid input column is monitored for each copy. If a fault affects the parity of the partitioning variables in one copy, a switch causes the output of the still valid copy to be used as the circuit output.

A problem arising with this design is that certain faults never affect the parity of partitioning variables in a currently valid input column and hence are undetectable. The authors claimed (page 656 of reference [2]) that faults affecting only nonpartitioning variables are "tolerable". This is, in fact, not true as is shown by the example given in Figure 3.

Suppose the y_1I_2 input to Y_4 becomes s-a-0 (we are assuming the use of shared logic to avoid problems observed in the previous section). This fault can never affect the parity of partitioning variables for a currently valid input column, because y_4 is a partitioning variable for I_3, and I_2 and I_3 are not adjacent input states. Hence this fault goes undetected.

But far from being tolerated, the fault has caused the circuit to lose its fault-secure property.

As evidence of this, suppose that the $\bar{x}_3$ input to AND gate y_1I_1 subsequently becomes s-a-1, causing the gate to realize $y_1(I_1 + I_2)$. Now if the circuit is stable in B under I_4 and the input changes to I_2, the stable state reached is B = 1010001 instead of the proper state C = 1001001. Hence the circuit has an incorrect output of 1 although no fault has been detected. In fact the circuit now realizes a flow table of the form shown in Figure 3(b).

Thus the example shows that designs relying on the monitoring of the controlling variables only may not, in general, be fault-secure.

IV. FAILSAFE ASYNCHRONOUS MACHINES

Mukai and Tohma have proposed a failsafe design for asynchronous sequential circuits [3]. We will show that this design suffers from problems discussed in the previous sections and the deficiency discussed below.

Nondetectable Faults

Figure 4 shows the minimum variable assignment and associated next state equations for the machine of Figure 3(a). In addition to the equations shown Mukai and Tohma propose the addition of a number of redundant terms to cause the machine to enter all 1's state in the presence of a s-a-1 fault affecting a controlling variable. They have demonstrated that a s-a-0 fault affecting a controlling variable will cause the all 0 state to be reached when the transition for which that controlling variable should be 1 is entered.

However, consider a s-a-0 fault on the y_1I_2 input to Y_4. Since y_4 is a controlling variable only for I_3, and I_2 and I_3 are nonadjacent this fault will never affect controlling variables. Thus it is not detectable under Mukai and Tohma's design. Now suppose the $\bar{x}_3$ input to the AND gate realizing y_1I_1 becomes s-a-1 causing the gate to realize $y_1(I_1 + I_2)$. This fault cannot cause any of the added redundant terms to excite because it does not affect the controlling variables of I_2. But when I_2 is entered from B the stable state reached will now be B with output 1 instead of the proper state C with output 0.

Thus even with the assumption that the redundant terms remain fault-free Mukai and Tohma's design is not fail-safe.

V. CONCLUSIONS

Several problems in the application of previously known results for the design of asynchronous circuits have been pointed out.

REFERENCES

1. D. H. Sawin and G. K. Maki, "Asynchronous Sequential Machines Designed for Fault Detection," IEEE Trans. Comp., C-23, pp. 239-249, March 1974.
2. D. H. Sawin and G. K. Maki, "Fault-Tolerant Asynchronous Sequential Machines," IEEE Trans. Comp., Vol. C-23, pp. 651-657, July 1974.
3. Y. Mukai and Y. Tohma, "A Method for the Realization of Fail-Safe Asynchronous Sequential Circuits," IEEE Trans. Comp., Vol. C-23, pp. 736-739, July 1974.
4. W. W. Patterson and G. Metze, "A Fail-Safe Asynchronous Sequential Machine," IEEE Trans. Comp., Vol. C-23, pp. 369-374, April 1974.
5. D. K. Pradhan and S. M. Reddy, "Fault-Tolerant Asynchronous Networks," IEEE Trans. Comp., Vol. C-22, pp. 662-668, July 1973.

						I_1	I_2	I_3	I_4
						x_1x_2			
y_1	y_2	y_3	y_4	y_5	y_6	00	01	10	11
1	0	1	0	1	0	Ⓐ	C	Ⓐ	D
0	1	0	1	0	1	-	D	A	Ⓑ
0	1	1	0	0	1	Ⓒ	Ⓒ	A	B
0	1	0	1	1	0	C	Ⓓ	A	Ⓓ

$$Y_1 = y_1I_1 + I_3$$

$$Y_2 = y_2I_1 + y_3I_2 + y_4I_2 + y_5I_4 + y_6I_4$$

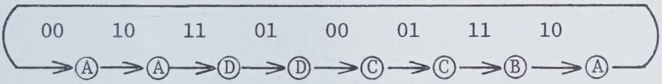
$$Y_3 = y_1I_1 + y_2I_1 + y_3I_2 + I_3$$

$$Y_4 = y_4I_2 + y_5I_4 + y_6I_4$$

$$Y_5 = y_1I_1 + y_4I_2 + I_3 + y_5I_4$$

$$Y_6 = y_2I_1 + y_3I_2 + y_6I_4$$

(a)



(b)

Figure 1. Example machine M_1

y_{Q_1}	y_{P_1}	y_{R_1}	y_{R_2}	I_Q	I_P	I_R
1 ...	1 ...	1	0	Ⓐ	B	D
. ...	1 ...	0	1	.	Ⓑ	G
⋮ ...	1 ...	0	1	⋮	B	G
	0 ...	1	0		F	Ⓓ
	⋮	⋮	⋮		⋮	⋮

$$Y_{Q_1} = y_{Q_1} I_Q + \ldots$$

$$Y_{P_1} = y_{Q_1} I_Q + y_{P_1} I_P + \ldots$$

$$Y_{R_1} = y_{Q_1} I_Q + y_{R_1} I_R + \ldots$$

$$Y_{R_2} = y_{P_1} I_P + y_{R_2} I_R + \ldots$$

Figure 2. Example machine M_2 (partial flow table shown)

							I_1	I_2	I_3	I_4	I_5
						$x_1x_2x_3$					
y_1	y_2	y_3	y_4	y_5	y_6	y_7	000	001	010	011	110
1	0	1	0	0	1	0	B	-	Ⓐ	D	C
1	0	1	0	0	0	1	Ⓑ /1	C	A	Ⓑ	-
1	0	0	1	0	0	1	B	Ⓒ /0	E	B	Ⓒ
0	1	0	0	1	1	0	Ⓓ	E	Ⓓ	Ⓓ	C
0	1	0	1	0	0	1	D	Ⓔ	Ⓔ	B	C

$$Y_1 = y_1I_1 + y_1I_2 + y_3I_3 + y_7I_4 + I_5$$

$$Y_2 = y_2I_1 + Y_2I_2 + y_4I_3 + y_5I_3 + y_6I_4$$

$$Y_3 = y_1I_1 + y_3I_3 + y_7I_4$$

$$Y_4 = y_1I_2 + y_2I_2 + y_4I_3 + I_5$$

$$Y_5 = y_2I_1 + y_5I_3 + y_6I_4$$

$$Y_6 = y_2I_1 + y_3I_3 + y_5I_3 + y_6I_4$$

$$Y_7 = y_1I_1 + y_1I_2 + y_2I_2 + y_4I_3 + y_7I_4 + I_5$$

Figure 3(a). Example machine M_3

I_1	I_2	I_3	I_4	I_5
B	-	Ⓐ	D	C
Ⓑ /1	Ⓑ /1	A	Ⓑ	-
B	B	E	B	Ⓒ
Ⓓ	E	Ⓓ	Ⓓ	C
D	Ⓔ	Ⓔ	B	C

Figure 3(b). Example machine M_3 with a fault

						000	001	010	011	110
y_1	y_2	y_3	y_4	y_5	y_6	I_1	I_2	I_3	I_4	I_5
1	0	1	0	1	0	B	–	Ⓐ	D	C
1	0	1	0	0	1	Ⓑ /1	C	A	Ⓑ	–
1	0	0	1	0	1	B	Ⓒ /0	E	B	Ⓒ
0	1	0	0	1	0	Ⓓ	E	Ⓓ	Ⓓ	C
0	1	0	1	0	1	D	Ⓔ	Ⓔ	B	C

$$Y_1 = y_1I_1 + y_1I_2 + y_3I_3 + y_6I_4 + (y_1 + y_2)\ I_5$$

$$Y_2 = y_2I_1 + y_2I_2 + y_2y_5I_3 + y_4I_3 + y_5I_4$$

$$Y_3 = y_1I_1 + y_3I_3 + y_6I_4$$

$$Y_4 = y_1I_2 + y_2I_2 + y_4I_3 + (y_1 + y_2)I_5$$

$$Y_5 = y_2I_1 + y_3I_3 + y_2y_5I_3 + y_5I_4$$

$$Y_6 = y_1I_1 + y_1I_2 + y_2I_2 + y_4I_3 + y_6I_4 + (y_1 + y_2)I_5$$

Figure 4. Example machine M_3 with minimum variable state assignment

PROPERTIES OF THE COMPLEX DIGITAL INVERSE FILTER AS APPLIED TO NARROWBAND INTERFERENCE REJECTION

FRANK W. SYMONS, JR.
Applied Research Laboratory
The Pennsylvania State University
State College, PA 16801

ABSTRACT

The properties of a complex digital inverse filter (DIF) are considered as used in the pre-whitening of narrowband interference prior to detection or parameter estimation based on LMS optimum filtering. The properties of output power, gain and frequency response are considered and it is shown how increased length of the filter has its usefulness in controlling the whitening of the input spectrum.

I. INTRODUCTION

Digital Inverse Filtering (DIF) has been widely investigated as a technique in speech processing [1-4]. Its useful property is that it uses the correlations between taps on an FIR (tapped delay line) filter structure to whiten the spectrum of an input speech process in an optimum (LMS) manner. In so doing, the information contained in the input process-corresponding to the non-white portions of the input spectrum-is determined as the inverse of the spectrum of the FIR which has been used to extract the colored (non-white) portion of the input process.

A second potentially useful application of the DIF is for a pre-whitening operation in a data communication system (sonar, radar) prior to detection or parameter estimation of a desired signal [5]. It is well known that most optimization schemes such as minimum mean square error (MMSE), maximum signal-to-noise ratio, and maximum gain require a pre-whitening operation. The DIF could be useful in performing the pre-whitening function-especially for the rejection of narrowband interference that does not overlap the spectrum of the desired signal.

This paper considers the properties of a complex digital inverse filter as used in the rejection of narrowband interference. It first derives (Section II) the optimum complex DIF and the corresponding expression for least mean square error (output power). The signal models to be used for detailed analysis are then derived (Section III) and the tap correlation matrix and its inverse derived. Optimum properties for the narrowband signal model are derived and discussed (Section IV) and conclusions drawn (Section V).

II. DERIVATION OF THE OPTIMUM FILTER

The form of the optimum filter is given by

$$A(z) = 1 + \sum_{i=1}^{M} a_i z^{-i} \tag{1}$$

where the a_i are complex weights on a tapped delay line of length M as shown in Figure 1. Tap signals x_i are assumed to be complex (quadrature) samples of the input process. The performance criterion, then, is to minimize the mean square error at the output of the filter by choosing the optimum set of complex weights $\{a_i\}$. In choosing this form of the filter, we are minimizing the mean square error between the summed outputs of the weighted taps and the input process. Theoretically, with a sufficiently long filter, all of the information (interference) on the tapped delay line that is correlated with the input process can be subtracted out. This leaves at the output only a process that is uncorrelated from point-to-point and the pre-whitening function is thus accomplished.

The optimum filter coefficients are derived by minimizing the mean square output of the filter which is given by

$$E\{|y|^2\} = E\{|x_o + \underline{a}^H \underline{x}|^2\} \quad , \tag{2}$$

$$= E\{(x_o + \underline{a}^H \underline{x})(x_o + \underline{a}^H \underline{x})^*\}$$

$$= E\{|x_o|^2 + \underline{a}^H \underline{x}\,\underline{x}^H \underline{a} + 2 \text{ Re } \underline{a}^H \underline{x}\, x_o^*\}$$

$$= E\{|x_o|^2\} + \underline{a}^H R_{xx} \underline{a} + 2 \text{ Re } \underline{a}^H \underline{p} \tag{3}$$

where $\underline{a}^T = [a_1 \; a_2 \; . \; . \; . \; a_M]$, $\underline{x}^T = [x_1 \; x_2 \; . \; . \; . \; x_M]$, R_{xx} is the tap correlation matrix = $E\{\underline{x}\,\underline{x}^H\}$ and $\underline{p}$ is the cross correlation vector (between the input signal x_o and the tap signals) $\underline{x}_o = E\{\underline{x}\, x_o^*\}$. The minimization is performed by completing the square of (3) in terms involving the complex weight vector $\underline{a}$. This results in the optimum weight vector

$$\underline{a} = -R_{xx}^{-1}\, \underline{p} \quad . \tag{4}$$

The optimum mean square output is obtained by substituting (4) into (3) to obtain

$$E\{|y|^2\} = E\{|x_o|^2\} - \underline{p}^H R_{xx}^{-1}\, \underline{p} \quad . \tag{5}$$

III. SIGNAL MODELS AND THE INVERSE CORRELATION MATRIX

The signals present at the input to the filter are assumed to be complex and narrowband. If the input to the filter is assumed to be the sum of L separate signals in the form

$$z_o(t) = \sum_{m=1}^{L} i_m(t) \tag{6}$$

then the waveform at the k^{th} tap is given by

$$z_k(t) = \sum_{m=1}^{L} i_m\,(t - k\,T_s) \quad . \tag{7}$$

In the case of an environment consisting of sources that are each narrowband compared to $f_s = \frac{1}{T_s}$, combined with an uncorrelated noise waveform, the pre-envelope of the waveform at the k^{th} tap is given by

$$x_k(t) = \sum_{m=1}^{L} i_m(t)\, e^{-jk\omega_m T_s} + n_k(t) \tag{8}$$

where $n_k(t)$ is the pre-envelope of the uncorrelated noise. The vector of tap waveforms is then

$$\underline{x}(t) = \begin{bmatrix} e^{-j\omega_1 T_s} & \cdot\cdot\cdot\cdot\cdot & e^{-j\omega_L T_s} \\ e^{-j\omega_1 (2T_s)} & & e^{-j\omega_L (2T_s)} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ e^{-j\omega_1 (MT_s)} & & e^{-j\omega_L (MT_s)} \end{bmatrix} \begin{bmatrix} i_1(t) \\ i_2(t) \\ \cdot \\ \cdot \\ \cdot \\ i_L(t) \end{bmatrix} + \begin{bmatrix} n_1(t) \\ n_2(t) \\ \cdot \\ \cdot \\ \cdot \\ n_M(t) \end{bmatrix} \quad . \tag{9}$$

If we now let $t = n\,T_s$ and define $\underline{x} \triangleq \underline{x}(t)\big|_{t = n\,T_s}$, $\underline{i} \triangleq i(t)\big|_{t = n\,T_s}$ and $\underline{n} = \underline{n}(t)\big|_{t = n\,T_s}$, (9) can be written as

$$\underline{x} \triangleq [\xi]\ \underline{i} + \underline{n} \quad . \tag{10}$$

The tap correlation matrix then becomes (assuming the signals to be uncorrelated with each other)

$$\begin{aligned} R_{xx} &= E\{\underline{x}\ \underline{x}^H\} \\ &= E\{([\xi]\ \underline{i} + \underline{n})([\xi]\ \underline{i} + \underline{n})^H\} \\ &= E\{[\xi]\ \underline{i}\ \underline{i}^H\ [\xi]^H\} + E\{\underline{n}\ \underline{n}^H\} \\ &\triangleq [\xi]\ R_{ii}\ [\xi]^H + R_{nn} \quad . \end{aligned} \tag{11}$$

The form of this matrix lends itself to analytical inversion by Woodbury's identity [5] as (defining R_{nn} as $\sigma_n^2\ I$ where I is the identity matrix and σ_n^2 is the noise variance on each tap)

$$R_{xx}^{-1} = \frac{1}{\sigma_n^2}\left\{I - [\xi]\ \frac{R_{ii}}{\sigma_n^2}\left[I + [\xi]^H\ [\xi]\ \frac{R_{ii}}{\sigma_n^2}\right]^{-1}\ [\xi]^H\right\} \quad . \tag{12}$$

In the case of a single colored interference, (12) becomes

$$\begin{aligned} R_{xx}^{-1} &= \frac{1}{\sigma_n^2}\left[I - \underline{\xi}\ \frac{\sigma_i^2}{\sigma_n^2}\left(1 + \underline{\xi}^H\ \underline{\xi}\ \frac{\sigma_i^2}{\sigma_n^2}\right)^{-1}\ \underline{\xi}^H\right] \\ &= \frac{1}{\sigma_n^2}\left[1 - \frac{\frac{\sigma_i^2}{\sigma_n^2}\ \ \underline{\xi}\ \underline{\xi}^H}{1 + \frac{\sigma_i^2}{\sigma_n^2}\ M}\right] \end{aligned} \tag{13}$$

where the matrix $[\xi]$ collapses to a single column $\underline{\xi}$ and σ_i^2 is the variance of the interference.

IV. OPTIMUM PROPERTIES

The results obtained in Section III are now applied to determine the optimum output power and gain at the frequency of interest and the frequency response of the optimum filter using the optimum weights derived.

A. Output Power

We want to evaluate the expression (5) to determine the optimum output power (MSE):

$$E\{|y|^2\} = E\{|x_o|^2\} + \underline{p}^H R_{xx}^{-1} \underline{p} \quad .$$

Consider each term:

(1)
$$\begin{aligned} E\{|x_o|^2\} &= E\{|\sum_{m=1}^{L} i_m + n_o|^2\} \\ &= E\{|\underline{1}^T \underline{i} + n_o|^2\} \\ &= E\{\underline{1}^T \underline{i}\,\underline{i}^H \underline{1} + |n_o|^2\} \\ &= \underline{1}^T R_{ii} \underline{1} + \sigma_n^2 \end{aligned} \tag{14}$$

where $\underline{1}^T \triangleq [1\ 1\ .\ .\ .\ 1]$.

Equation (14) collapses to

$$E\{|x_o|^2\} = \sigma_i^2 + \sigma_n^2 \tag{15}$$

for a single interference.

(2) $\underline{p}^H R_{xx}^{-1} \underline{p}$

We know

$$\begin{aligned} \underline{p} &= E\{\underline{x}\, x_o^*\} \\ &= E\{([\xi]\, \underline{i} + \underline{n})\, x_o^*\} \\ &= E\{([\xi]\, \underline{i} + \underline{n})(\underline{1}^T \underline{i} + n_o)^*\} \\ &= E\{[\xi]\, \underline{i}\,\underline{i}^H \underline{1}\} \\ &= [\xi]\, R_{ii}\, \underline{1} \quad . \end{aligned} \tag{16}$$

Equation (16) collapses to

$$\underline{p} = \sigma_i^2\, \underline{\xi} \quad . \tag{17}$$

Then using (13) and (16):

$$\underline{p}^H R_{xx}^{-1} \underline{p} = \frac{\sigma_i^2}{\sigma_n^2} \underline{\xi}^H \quad 1 - \frac{\frac{\sigma_i^2}{\sigma_n^2} \underline{\xi}\, \underline{\xi}^H}{1 + \frac{\sigma_i^2}{\sigma_n^2} M} \quad \underline{\xi}\, \sigma_i^2 = \sigma_i^2 \; \frac{\frac{\sigma_i^2}{\sigma_n^2} M}{1 + \frac{\sigma_i^2}{\sigma_n^2} M} \quad . \tag{18}$$

Substituting (15) and (18) into (5) gives the optimum output power

$$E\{|y|^2\} = \sigma_i^2 + \sigma_n^2 - \sigma_i^2 \; \frac{\frac{\sigma_i^2}{\sigma_n^2} M}{1 + \frac{\sigma_i^2}{\sigma_n^2} M}$$

$$= \sigma_n^2 + \frac{\sigma_i^2}{1 + \frac{\sigma_i^2}{\sigma_n^2} M} \quad . \tag{19}$$

B. Optimum Gain

The power gain of the filter at the frequency of the interference then becomes:

$$\text{GAIN} = \frac{P_{out}}{P_{in}}$$

$$= \frac{E\{|y|^2\}}{\sigma_i^2 + \sigma_n^2} \tag{20}$$

$$= \frac{\sigma_n^2 + \frac{\sigma_i^2}{1 + \frac{\sigma_i^2}{\sigma_n^2} M}}{\sigma_i^2 + \sigma_n^2}$$

$$= \frac{1 + \frac{\sigma_i^2}{\sigma_n^2} (M + 1)}{(1 + \frac{\sigma_i^2}{\sigma_n^2})(1 + \frac{\sigma_i^2}{\sigma_n^2} M)} \quad . \tag{21}$$

C. Effects of the Parameters on Optimum Results

There are two parameters of interest here: $\frac{\sigma_i^2}{\sigma_n^2}$ and M. The limiting cases would be for the $\frac{\sigma_i^2}{\sigma_n^2} M$ product to be small and large.

(1) $\frac{\sigma_i^2}{\sigma_n^2} M$ small.

In this case, the optimum results become:

$$P_{out} = \sigma_n^2$$

and

$$\text{Gain} = 1 \quad .$$

These values indicate a white noise situation which is exactly what the filter is to accomplish. If the input is white noise, the filter should pass it without distortion.

(2) $\frac{\sigma_i^2}{\sigma_n^2} M$ large.

In this case, the results become:

$$P_{out} = \frac{M+1}{M} \sigma_n^2$$

and

$$\text{Gain} = \frac{M+1}{M} \frac{\sigma_n^2}{\sigma_i^2}$$

It can be seen that the results are relatively insensitive to the filter order but for reasonable length the output power is purely noise power and the gain depends primarily on the interference-to-noise ratio for large $\frac{\sigma_i^2}{\sigma_n^2}$. The effect of filter order is significant in terms of the frequency response as will be seen in the next section.

D. Frequency Response

We want to obtain the frequency response of the whitener for a single interference. The filter coefficients are assumed to be designed for an arbitrary frequency ω_o. The input interference will be allowed to vary over a range of frequencies while maintaining the same power.

The transfer function, $G(\omega)$, is given by

$$G(\omega) = \frac{P_{out}(\omega)}{P_{in}(\omega)} \quad . \tag{22}$$

Let the correlation matrix for the input source be

$$R(\omega) = \sigma_n^2 I + \sigma_i^2 \underline{\xi}(\omega) \underline{\xi}^H(\omega) \triangleq \sigma_n^2 I + \sigma_i^2 \underline{\xi}_\omega \underline{\xi}_\omega^H \quad . \tag{23}$$

The output power at arbitrary frequency is then

$$P_{out} = E\{|x_o|^2\} + a^H R(\omega) \underline{a} - 2 \operatorname{Re} \underline{a}^H \underline{p}(\omega) \quad . \tag{24}$$

But

$$\underline{a} = -R_{xx}^{-1}\,\underline{p} = -\,[\sigma_n^2 I + \sigma_i^2 \underline{\xi}\;\underline{\xi}^H]^{-1}\;[\sigma_i^2 \underline{\xi}] = \frac{\dfrac{\sigma_i^2}{\sigma_n^2}}{1 + \dfrac{\sigma_i^2}{\sigma_n^2} M}\;\underline{\xi}\;. \tag{25}$$

Since $P_{in}(\omega) = \sigma_n^2 + \sigma_i^2$ for all ω, the transfer function becomes

$$G(\omega) = \frac{P_{out}(\omega)}{\sigma_n^2 + \sigma_i^2}\;. \tag{26}$$

Combining Equations (23), (24), and (25), (26) can be manipulated into the form

$$G(\omega) = c_1 - c_2 \cos\left[\frac{M+1}{2}(\omega_o - \omega)T_s\right] \frac{\sin[\frac{M}{2}(\omega_o - \omega)T_s]}{\sin[\frac{1}{2}(\omega_o - \omega)T_s]}$$

$$+\; c_3 \frac{\sin^2[\frac{M}{2}(\omega_o - \omega)T_s]}{\sin^2[\frac{1}{2}(\omega_o - \omega)T_s]}\;, \tag{27}$$

where

$$c_1 = 1 + \frac{M}{1 + \sigma_i^2/\sigma_n^2}\,b^2\;,$$

$$c_2 = 2b\,\frac{\sigma_i^2/\sigma_n^2}{1 + \sigma_i^2/\sigma_n^2}\;,$$

$$c_3 = \frac{\sigma_i^2/\sigma_n^2}{1 + \sigma_i^2/\sigma_n^2}\,b^2$$

and

$$b = \frac{\sigma_i^2/\sigma_n^2}{1 + \dfrac{\sigma_i^2}{\sigma_n^2} M}\;. \tag{28}$$

Frequency Response Limits. Again we want to allow $\frac{\sigma_i^2}{\sigma_n^2}$ to vary.

(1) $\frac{\sigma_i^2}{\sigma_n^2}$ small.

In this case, $c_1 \to 1$, $c_2 \to 0$, $c_3 \to 0$, and $G(\omega) \to 1$ as would be the case if only white noise were present.

(2) $\frac{\sigma_i^2}{\sigma_n^2}$ large.

In this case, for ω not 'close to' ω_o, $c_1 \to 1$, $c_2 \to \frac{2}{M}$, $c_3 \to \frac{1}{M^2}$. As M gets large, c_2 and c_3 go to zero, leaving $G(\omega) = 1$. Thus, as M gets larger the frequency response approaches a flat spectrum. Of course, for ω 'close to' ω_o, the frequency response drops to its optimum level. 'Close to' means within the 'main lobe' of $\sin \Theta/\Theta$ function. As M gets larger, the frequency band of the main lobe gets smaller, and the 'notch' created by the filter gets narrower with the frequency response of the rest of the band moving toward one.

Performance curves illustrating this point are shown in Figure 2 for $\sigma_i^2 = 100$, $\sigma_n^2 = 1$, $M = 5$, and $M = 40$.

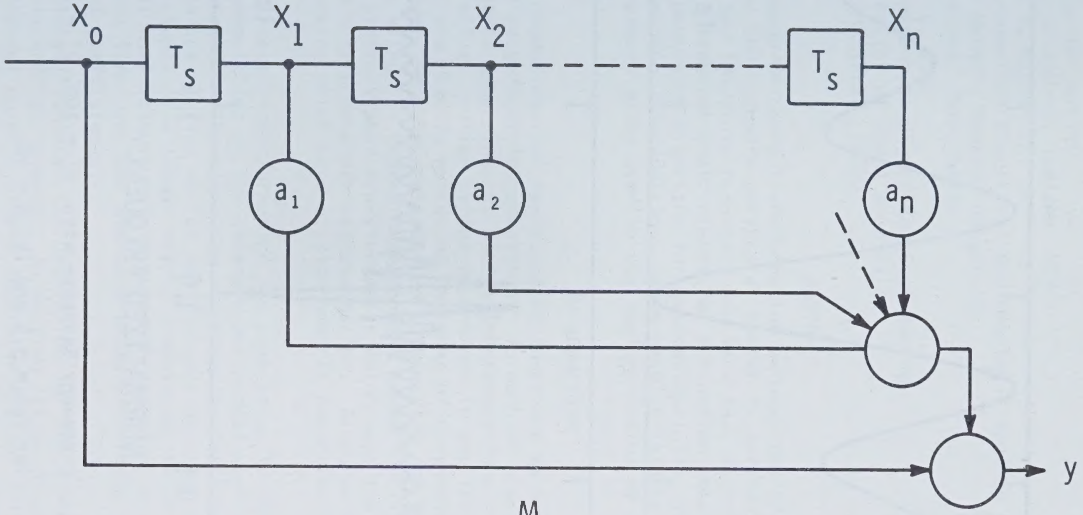
V. DISCUSSION AND CONCLUSIONS

The optimum DIF was derived for the complex data case. Analysis of the narrowband case shows its performance to be dependent on the interference power to uncorrelated noise power ratio and the length of the filter. The frequency response analysis shows the same parameter dependence. It is shown that the higher the interference-to-noise ratio, the deeper the notch that is formed and the higher the filter order, the narrower the notch that is formed and the flatter the spectrum over the rest of the band.

This application of the DIF appears to have significant potential. The key to its success lies in determining a pragmatic way to implement the filter in real time.

REFERENCES

1. Atal, A. S. and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acous. Soc. Am., Vol. 50, No. 2, Pt. 2, 1971, pp. 637-655.
2. Markel, J. B., "Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation," IEEE Trans. on Aud. and Electroacoustics, Au-20, No. 2, June 1972, pp. 129-137.
3. Markel, J. B. and A. H. Gray, "On Autocorrelation Equations as Applied to Speech Analysis," IEEE Trans. on Audio and Electroacoustics, Au-21, No. 2, April 1973, pp. 69-79.
4. Wakita, H., "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms," IEEE Trans. on Audio and Electroacoustics, Au-21, No. 5, October 1973, pp. 417-427.
5. Symons, F. W., Jr., "Spatial and Spectral Filtering as Applied to Array Processing," Ph.D. Dissertation, The Pennsylvania State University, University Park, PA, 1975.


X0
Ts
X1
Ts
X2
Ts
Xn
a1
a2
an
y

$$A(Z) = 1 + \sum_{i=1}^{M} a_i Z^{-i}$$

Figure 1. Structure of the Optimum Filter

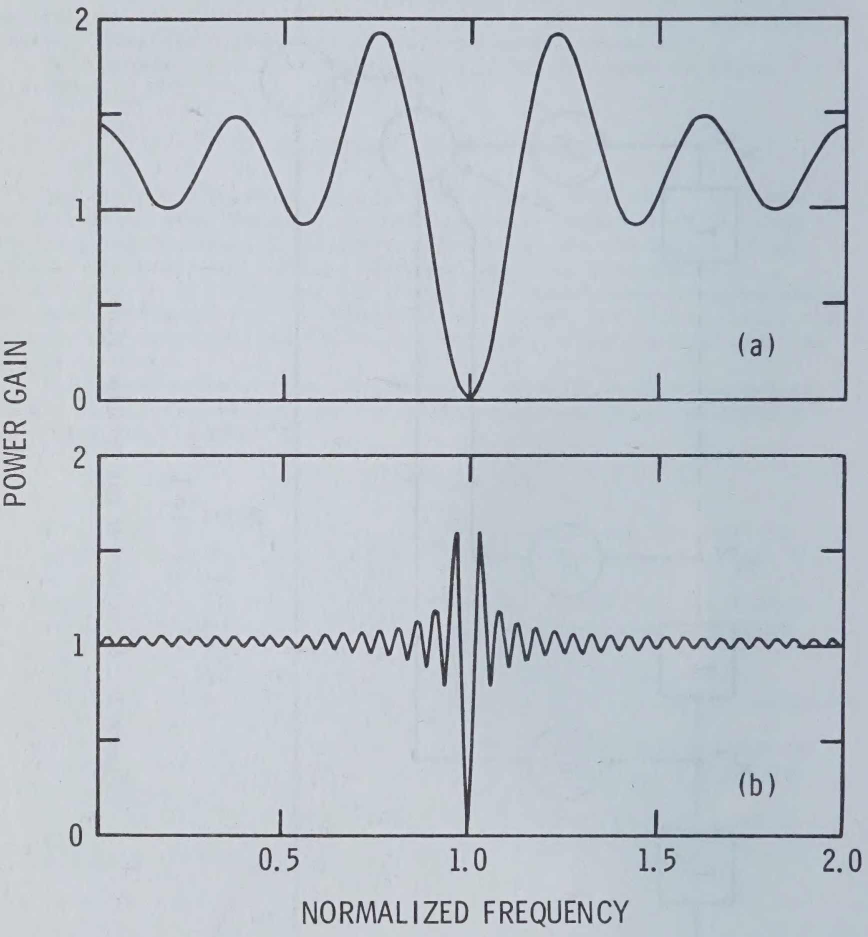


Figure 2. Frequency Response with $\sigma_i^2 = 100$, $\sigma_n^2 = 1$ for (a) $M = 5$ and (b) $M = 40$

FOURIER SERIES REPRESENTATION FOR POLYNOMIALS WITH APPLICATION TO NONLINEAR DIGITAL FILTERING

NEAL C. GALLAGHER, JR.
School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907
JOHN W. ALLEN
Department of Electrical Engineering & Applied Physics
Case Western Reserve University
Cleveland, Ohio 44106
GARY L. WISE
Department of Electrical Engineering
University of Texas at Austin
Austin, Texas 78712

ABSTRACT

A change-of-basis transformation between the trigonometric Fourier Series and the Legendre polynomial series is derived. This transformation leads to two immediate results: one is a fast algorithm for computing the Legendre polynomial coefficients for a function; the second is a numerically stable computation algorithm for evaluating $j_n(z)$, the n^{th} order spherical Bessel function. In addition, these results lead to a simple recursive digital network to be used in the implementation of adaptive nonlinear functions.

I. INTRODUCTION

Two common ways of representing functions have been polynomial expansions and trigonometric expansions. In much of electrical engineering the trigonometric expansion has useful interpretations and has dominated over the generalized Fourier series expansions in applications. However, many functions are readily expressed in terms of polynomials. For example, in a recent paper, we have presented an interesting representation for bandlimited functions using the Legendre polynomial expansion [1]. In this paper we derive a simple linear transformation which maps the polynomial representation into a trigonometric representation. Also, we derive the inverse transformation which maps a trigonometric expansion to a polynomial expansion.

The inverse transformation has enabled us to develop a fast algorithm for the computation of the Legendre polynomial coefficients for any L_2 $[-\pi,\pi]$ function. The algorithm utilizes the Fast Fourier Transform (FFT) to compute the Fourier series coefficients and then multiplies the vector of coefficients by a linear matrix transformation to compute the vector of polynomial coefficients. This approach can offer a considerable saving in computation time over the standard integral formula for computing these polynomial coefficients.

Section II contains the derivation for the elements of the transformation matrix and its inverse. In section III we discuss how these results lead to the design of a simple recursive digital filter that may be employed to implement almost any zero memory non-linear operation (ZNL). In section IV we discuss several computational considerations, and in section V we present an example.

II. DEVELOPMENT

Assume $H(x)$ is a polynomial defined on $[-\pi,\pi]$. By writing $H(x)$ in Legendre polynomials, we can use results developed in an earlier paper to easily write the Fourier series expansion for $H(x)$ that converges in $[-\pi,\pi]$.

Assume that

$$H(x) = \sum_{n=0}^{N} a_n P_n(\tfrac{x}{\pi}), \qquad -\pi \leq x \leq \pi \tag{1}$$

where $P_n(\cdot)$ is the n^{th} Legendre polynomial. To determine the Fourier series representation for $H(x)$, it is sufficient to know the Fourier series expansion for $P_n(\frac{x}{\pi})$. From [1] we know that

$$P_n(\tfrac{x}{\pi}) = \sum_{m=-\infty}^{\infty} [(i)^n j_n(m\pi)] e^{-imx} . \tag{2}$$

By Eqs. (1) and (2), we can write

$$H(x) = \sum_{n=-\infty}^{\infty} h_n e^{-inx} , \tag{3a}$$

where

$$h_n = \sum_{m=0}^{N} a_m (i)^m j_m(n\pi) . \tag{3b}$$

We note that the sequence $\{h_n\}$ can be treated as a matrix multiplied by a vector (in an infinite dimensional vector space). That is,

$$h = a B,$$

where a is a row vector of polynomial coefficients

$$a = [a_o, a_1, \ldots, a_N]$$

and B is the $(N+1) \times \infty$ transformation matrix whose $(m,n)^{th}$ element is given by

$$b_{mn} = (i)^m j_m(n\pi) .$$

Observe that the transformation matrix B is independent of the function $H(x)$, and therefore a different transformation matrix need not be computed for each different $H(x)$. A procedure for the computation of the terms $j_m(n\pi)$ is contained in section IV.

Now we consider the case where $H(x)$ is not a polynomial. Assume that $H(x)$ belongs to $L_2[-\pi,\pi]$, and therefore possesses a Legendre polynomial series expansion convergent in $L_2[-\pi,\pi]$. The coefficient vector becomes infinite dimensional, and the transformation matrix B is doubly infinite. As a result, Eq. (3b) is written as

$$h_n = \sum_{m=0}^{\infty} a_m (i)^m j_m(n\pi). \tag{4}$$

It can be shown [1] that the sum in Eq. (4) converges uniformly in n. That is, for any $\varepsilon > 0$, there exists a K such that for all $M \geq K$

$$|h_n - h_n(M)| < \varepsilon$$

for all n, where

$$h_n(M) = \sum_{m=0}^{M} a_m (i)^m j_m(n\pi).$$

Thus the sequence $\{h_n(M)\}$ is a uniform approximation to the sequence of Fourier series coefficients for the function $H(x)$.

Now, we consider the inverse transformation denoted by B^{-1}. The computation of B^{-1} directly from the matrix B can be performed; however, it is best to compute the elements of B^{-1} in a fashion similar to the computation of the elements of B. First, we note that the coefficients $\{a_n\}$ in Eq. (1) are given by the relation [1]

$$a_n = \frac{2n+1}{2\pi} \int_{-\pi}^{\pi} H(x)\, P_n(\tfrac{x}{\pi})\, dx \,. \tag{6}$$

Using the Fourier series expansion for H(x) found in Eq. (3a), we find

$$a_n = (2n+1) \sum_{m=-\infty}^{\infty} h_m \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} P_n(\tfrac{x}{\pi})\, e^{-imx}\, dx \right] . \tag{7}$$

Arguing as in the derivation of Eq. (2), we have

$$a_n = (2n+1) \sum_{m=-\infty}^{\infty} h_m [(-i)^n\, j_n(m\pi)] \,. \tag{8}$$

If we define

$$\hat{b}_{mn} = (2n+1)\ (-i)^n\, j_n(m\pi) \,, \tag{9a}$$

then

$$[\hat{b}_{mn}] = B^{-1} \,; \tag{9b}$$

that is, the matrix elements of B^{-1} are given by $\hat{b}_{mn}$. Consequently, Eq. (8) may be rewritten as

$$a = h\, B^{-1}, \tag{10}$$

where h is the row vector of Fourier series coefficients and a is the row vector of polynomial coefficients. In practice, a finite number of elements of h are computed by use of the FFT algorithm; we then perform the vector multiplication indicated by Eq. (10) to compute a finite number of elements of the vector a. This procedure for computing the Legendre polynomial coefficients for a function can provide a great reduction in computation time as compared to the direct evaluation of the integral in Eq. (6).

In the next section, we discuss a technique for the implementation of nonlinear operations by use of a simple recursive digital structure.

III. DIGITAL ZNL IMPLEMENTATION

We begin with the following recursion relation for Legendre Polynomials

$$(n+1)\, P_{n+1}(x) = (2n+1)\, x\, P_n(x) - n\, P_{n-1}(x) \,, \tag{11}$$

where

$$P_o(x) = 1$$

$$P_1(x) = x \,, \qquad -1 \le x \le 1 \,.$$

It is straight forward to implement this relation with a recursive digital network. Then, given this method by which to generate $\{P_n(x)\}$ for a given x, and given the coefficients $\{a_n\}$, the nonlinear function H(x) is easily constructed as

$$H(\pi x) \approx \sum_{n=0}^{N} a_n\, P_n(x), \quad -1 \le x \le 1. \tag{12}$$

Figure 1 contains one possible implementation of Eqs. (11) and (12). In Fig. 1 a recursive network is used to generate $P_n(x)$ for successive values of n. With the coefficients $\{a_n\}$ stored in memory, the products and sum indicated by Eq. (12) are then performed. We note that the function H(x) implemented by this network may be easily changed simply by changing the coefficients $\{a_n\}$; consequently, except for a memory unit, the network remains unchanged regardless of the function H(x).

Suppose that in order to obtain acceptable accuracy at the system output it is necessary, to make N = 50 in Eq. (12). Thus, given a value x we must wait for 50 terms to be accumulated before we have the output H(x). Therefore this digital ZNL could reduce the system data rate by a factor of 50; however, for a price, it is possible to eliminate this rate reduction. One approach would be to simply use a faster clock rate with the ZNL structure. Another approach is to construct N = 50 such ZNL structures; the system's first input sample, call it x_1, goes into the first ZNL; the second sample, x_2, goes into the second ZNL and so on. Now, just as x_{50} is going into the 50th ZNL, we are getting the output $H(x_1)$ from the first ZNL. Consequently, after an initial delay of 50 samples, 50 separate parallel networks can operate at a rate 50 times greater than a single network. Clearly, there are compromises that can be made between hardware cost and data rate.

Potential applications of this ZNL structure include adaptive nonlinear systems. In such an application the coefficients $\{a_n\}$ can be continually updated by use of the fast algorithms previously discussed.

IV. COMPUTATIONAL CONSIDERATIONS

In this section we discuss two different methods for the computation of $j_n(m\pi)$. We begin by noting the following relation for generating values of $j_n(z)$ [2]

$$j_n(z) = f_n(z) \sin z + (-1)^{n+1} f_{-n-1}(z) \cos z, \tag{13}$$

where

$$f_o(z) = \frac{1}{z}, \qquad f_1(z) = z^{-2}$$

and

$$f_{n-1}(z) + f_{n+1}(z) = (2n+1)\, z^{-1} f_n(z) .$$

Consequently,

$$j_n(m\pi) = (-1)^{n+1+m} f_{-n-1}(m\pi), \qquad n = 1,2,\ldots$$

where

$$j_o(m\pi) = \begin{cases} 1, & m = 0 \\ 0, & \text{otherwise} \end{cases} .$$

The recursion in Eq. (13) appears to provide a relatively simple method for computing $j_n(m\pi)$; unfortunatly, this relation is numerically unstable. We have found that even with double precision arithmetic error accumulates rapidly. When using this recursion relation, we have produced overflow conditions for n = 20 and $z = \pi$, 2π on a UNIVAC 1108 digital computer; this is, of course, unacceptable.

Fortunately, we have developed a numerical technique that appears to be very stable. First, we note that the recursion relation for generating $P_n(x)$ in Eq. (11) is very tolerant of roundoff accumulation. Then, we observe that Eq. (2) implies the following approximate relationship.

$$P_n\left(\frac{2q}{N}\right) = \sum_{m=-\frac{N}{2}+1}^{N/2} \left[(i)^n\, j_n(m\pi)\right] e^{-i\frac{2\pi mq}{N}} \qquad q = -\frac{N}{2}+1,\ldots,\frac{N}{2} \tag{14}$$

which is the discrete Fourier transform relation. Consequently, we have the inverse relationship

$$(i)^n\, j_n(m\pi) = \frac{1}{N} \sum_{q=-\frac{N}{2}+1}^{N/2} P_n\left(\frac{2q}{N}\right) e^{i\frac{2\pi mq}{N}}, \qquad m = -\frac{N}{2}+1,\ldots,\frac{N}{2} \tag{15}$$

Except for aliasing error, Eq. (15) is accurate. In order to reduce aliasing error it is necessary to increase N; generally speaking, N must be chosen separately for each case, depending on the degree of the polynomial $P_n(\cdot)$. Our procedure for computing $j_n(m\pi)$ is as follows: using Eq. (11), values of $P_n(\frac{2q}{N})$ are generated; then, using Eq. (15), values of $j_n(m\pi)$ are computed. The values of $j_n(m\pi)$ used in the following section's example have been computed by this method. We now discuss factors considered in the computation of the coefficients $\{a_n\}$.

If we assume that H(x) is real-valued, then Eq. (8), used to evaluate a_n, can be simplified by noting that the Fourier coefficients $\{h_m\}$ satisfy the relation $h_{-m} = h_m^*$. Also, $j_n(m\pi)$ satisfies the relation [2]

$$j_n(-m\pi) = (-1)^n\, j_n(m\pi)\ . \tag{16}$$

Consequently, Eq. (8) may be rewritten as

$$\begin{aligned} a_o &= h_o \\ a_n &= 2(2n+1)(-1)^{\frac{n}{2}} \sum_{m=1}^{\hat{N}/2} j_n(m\pi)\ \mathrm{Re}\ \{h_m\}\ , \text{ n-even and } \neq 0 \\ a_n &= 2(2n+1)(-1)^{\frac{n-1}{2}} \sum_{m=1}^{\hat{N}/2} j_n(m\pi)\ \mathrm{Im}\ \{h_m\}\ , \text{ n-odd,} \end{aligned} \tag{17}$$

where Re $\{h_m\}$ and Im $\{h_m\}$ denote the real and imaginary parts of h_m, respectively; and the maximum value for $\hat{N}$ is N, which is the number of coefficients $\{h_m\}$ that are computed. By use of Eq. (17), the polynomial coefficients $\{a_n\}$ are computed from the Fourier coefficients $\{h_n\}$.

The following section contains a numerical example computed with the aid of the results summarized in Eqs. (15) and (17).

V. AN EXAMPLE

In this section we will demonstrate the computation of the polynomial coefficients $\{a_n\}$ for

$$\begin{aligned} H(\pi x) &= P_3(x) + P_2(x) \\ &= \frac{1}{2}\,(5x^3 + 3x^2 - 3x - 1), \quad -1 \le x \le 1\ . \end{aligned}$$

By comparison with Eq. (1), we see that all $a_n = 0$ with the exception of $a_2 = a_3 = 1$. For this example, all computations are single precision arithmetic. The Fourier coefficients $\{h_m\}$ are computed by use of the DFT relation:

$$h_m = \frac{1}{N} \sum_{p=-\frac{N}{2}+1}^{N/2} H\left(\frac{2\pi p}{N}\right) e^{i\frac{2\pi mp}{N}} , \quad m = -\frac{N}{2}+1,\ldots,\frac{N}{2} ,$$

and the terms $\{j_n(m\pi)\}$ are computed by use of Eq. (15). Finally, the coefficients $\{a_n\}$ are computed by application of Eq. (17). These computations are performed for three different values of $N = \hat{N}$: $N = \hat{N} = 8192$, 4096, and 2048. Table 1 contains typical values of $\{a_n\}$.

N	2048	4096	8192	True Value
a_o	-4.880×10^{-4}	-2.441×10^{-4}	-1.220×10^{-4}	0
a_1	-1.463×10^{-3}	-7.320×10^{-4}	-3.661×10^{-4}	0
a_2	0.9976	0.9988	0.9994	1
a_3	0.9966	0.9983	0.9991	1
a_4	-4.385×10^{-3}	-2.195×10^{-3}	-1.098×10^{-3}	0
a_5	-5.353×10^{-3}	-2.681×10^{-3}	-1.342×10^{-3}	0

Table 1. Selected Values of $\{a_n\}$ Computer for $N = \hat{N}$.

As the value of N is increased the effect of aliasing is reduced in our computation of $\{j_n(m\pi)\}$ and $\{h_m\}$. Also, as the value of $\hat{N}$ is increased, the summation terms of Eq. (17) become more accurate. Inspection of Table 1 reveals that as the value of $N = \hat{N}$ is doubled, the error in computation is halved. We ask, to which parameter, N or $\hat{N}$, is the error most sensitive? Some insight may be obtained by examining the value of a_o in Table 1; the value of a_o is independent of the value of $\hat{N}$ (see Eq. 17). Because the computation error for a_o is halved as N doubles, the implication is that much of the error is due to aliasing and is dependent on N. This is illustrated in Table 2, where the parameter $\hat{N}$ is held constant at $\hat{N} = 2048$.

N	2048	4096	8192	True Value
a_o	-4.880×10^{-4}	-2.441×10^{-4}	-1.220×10^{-4}	0
a_1	-1.463×10^{-3}	-8.322×10^{-4}	-6.545×10^{-4}	0
a_2	0.9976	0.9988	0.9994	1
a_3	0.9966	0.9981	0.9985	1
a_4	-4.385×10^{-3}	-2.192×10^{-3}	-1.096×10^{-3}	0
a_5	-5.353×10^{-3}	-3.049×10^{-3}	-2.399×10^{-3}	0

Table 2. Selected Values of $\{a_n\}$ for $\hat{N} = 2048$.

First note that in Tables 1 and 2, the values of $\{a_n\}$ for $N = 2048$ are identical because the same value of $\hat{N} = 2048$ is used in both cases. Upon comparing the two tables, we note that the value of $\hat{N}$ does affect the computation error; for some value $\{a_n\}$ the effect is negligable while with others a factor of 2 difference is noted.

The results of this section indicate the importance of computing the elements of the B^{-1} matrix very precisely, specifically $\{j_n(m\pi)\}$. If this computation is accomplished with the DFT procedure outlined previously, then a large value of N should be chosen for the DFT; we emphasize that this computation need be made only once, and the results stored in memory.

V. DISCUSSION AND SUMMARY

In this paper, we have developed a fast algorithm for the computation of the Legendre polynomial coefficients of a function. Combining this fast algorithm with a digital implementation of the Legendre polynomial recursion relation, it is possible to implement an adaptive nonlinearity. Generally, the more complicated the non-linearity to be implemented, the higher will be the degree of the polynomial approximation required; this in turn, reduces the throughput rate of the non-linear device unless we are willing to pay a greater price in hardware. Nevertheless, by using this approach one acquires a large degree of flexibility in specifying the characteristics of the nonlinearity.

In the past, most non-linear devices used in practice have had very simple characteristics; in part, this has been due to the relative difficulty in implementing more complicated devices. However, as signal processing schemes become increasingly sophisticated, the need for complex or adaptive non-linear devices is growing. One very interesting application is in robust or distribution free signal detection schemes [3] where an adaptive, easily implmentable, non-linearity can be of great benefit.

In addition to its application to non-linear systems, we should not overlook the advantage of having a fast algorithm for the computation of the polynomial expansion for almost any function.

It should also be noted that there are many special function pairs with similar Fourier transform relationships that may be exploited in a fashion similar to that presented in this paper [4]. The fact that there may not be a simple expression for computing B^{-1} should not be a deterent from performing the computations because B^{-1} need only be computed once and then stored in memory for all future requirements.

ACKNOWLEDGEMENT

The support of the National Science Foundation under Grant ENG 76-09443, and of the Air Force Office of Scientific Research under Grant AFOSR-76-3062 is gratefully acknowledged.

REFERENCES

1. N. C. Gallagher and G. L. Wise, "A Representation for Band-Limited Functions," Proc. of the IEEE, Vol. 63, pp. 1624-1625, Nov. 1975.

2. M. Abromowitz and I. A. Segun, Handbook of Mathematical Functions, Dover Publications, New York, N.Y., p. 782.

3. S. A. Kassam and S. B. Thomas, "Generalizations of the Sign Detector Based on Conditional Tests," IEEE Trans. on Communications, Vol. COM-24, pp. 481-487, May 1976.

4. K. Yao and J. B. Thomas, "On Band-Limited Properties of Fourier Transform Pairs of Some Special Functions," Proc. 3rd Annu. Allerton Conf. on Circuit and System Theory, pp. 299-309, Oct. 1965.

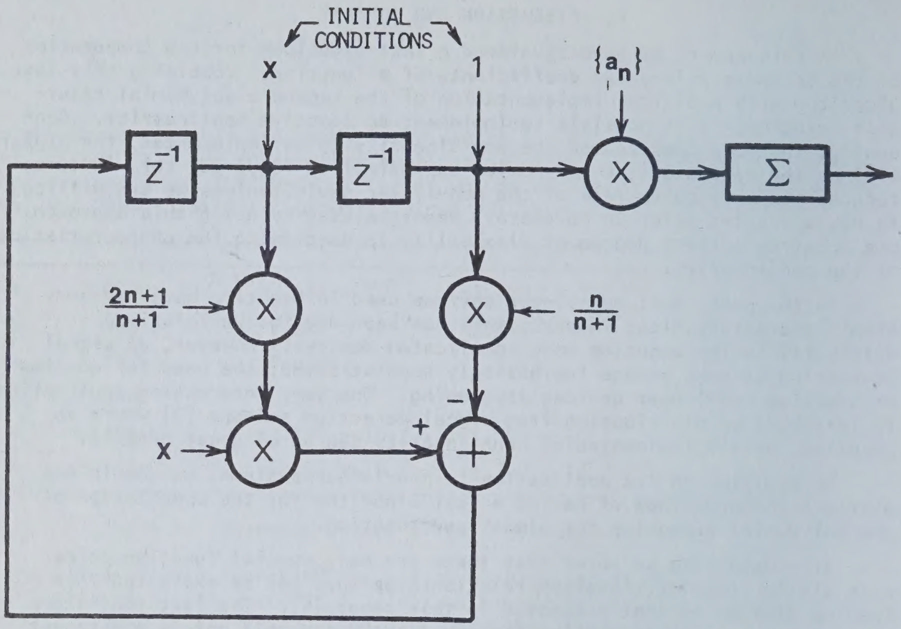

INITIAL CONDITIONS
X
1
{a_n}
Z^-1
Z^-1
X
Σ
2n+1 / n+1
X
X
n / n+1
−
+
X
X
+

Fig. 1. Digital Network for the Computation of $\Sigma\ a_n\ P_n(x)$.

STEADY-STATE, ROUNDOFF-ERROR ANALYSIS IN DIGITAL FILTERS REALIZED WITH FIXED-POINT ARITHMETIC†

A. SRIPAD and D. L. SNYDER
Washington University, St. Louis, Missouri 63130

SUMMARY

Finite word-length in the implementation of a filter with digital devices introduces unavoidable quantization errors. The effect of such quantization in the degradation of performance of the filter compared with its infinite precision realization has gained considerable attention. We can identify the quantization error in the fixed-point digital-filter implementation as due to input, coefficient, and product quantizations. In this paper, we are concerned only with the roundoff error due to finite-precision multiplication, with no overflow in addition operations. Consider a constant system represented by

$$x_{k+1} = Ax_k + Bu_k, \quad y_k = Cx_k + Du_k \tag{1}$$

where $u_k \in R^m$, $y_k \in R^r$, $x_k \in R^n$ are the input, output and state vectors. A, B, C and D are matrices. Each non-integer multiply in (1) is modelled as an infinite precision multiply and an additive, white noise of zero mean and variance $2^{-2b}/12$, where b bits represent the fractional number. Define

$$\begin{aligned} P &\triangleq (2^{-2b}/12)\,\mathrm{diag}\,(p_1, p_2, \ldots, p_n) \\ Q &\triangleq (2^{-2b}/12)\,\mathrm{diag}\,(q_1, q_2, \ldots, q_r) \end{aligned} \tag{2}$$

where p_i and q_i represent the total number of non-integer entries in the *ith* row of A, B and C, D respectively. Then the set of stochastic difference equations driven by white noise sources to represent state and output roundoff errors (ε_k and z_k) are

$$\begin{aligned} \varepsilon_{k+1} &= A\,\varepsilon_k + w_k, \quad \varepsilon_0 = 0 \\ z_k &= C\varepsilon_k + v_k \end{aligned} \tag{3}$$

where $E(w_k)=0$, $E(v_k)=0$, $E(w_k w_j')=P\delta_{kj}$, $E(v_k v_j')=Q\delta_{kj}$, and $E(w_k v_j')=0$. For asymptotically stable A, we get the steady-state error means to be zero and the covariances to be

$$\lim_{k\to\infty} E[\varepsilon_k\, \varepsilon_k'] = \Phi = A\Phi A' + P \tag{4}$$

$$\lim_{k\to\infty} E[z_k\, z_k'] = \Psi = C\Phi C' + Q. \tag{5}$$

Eq. 4 is the matrix Lyapunov equation, which has been widely studied in the context of stability theory. Now consider a non-singular transformation, $\xi_k = T^{-1}x_k$, to change the system coordinates; then (1) becomes,

$$\xi_{k+1} = T^{-1}AT\xi_k + T^{-1}B\, u_k, \quad y_k = CT\xi_k + Du_k. \tag{6}$$

In the infinite precision implementation, (1) and (6) would have the same system response, but with finite precision arithmetic, the output error covariance depends on the coordinate system. We compare the error covariances of two systems by

$$\Psi - \Psi_T = C(\Phi - T\Phi_T T')C' + (Q - Q_T) \tag{7}$$

$$(\Phi - T\Phi_T T') = A(\Phi - T\Phi_T T')A' + (P - TP_T T'), \tag{8}$$

where the subscript T denotes the quantities for the transformed system. Eq. (8) is also a matrix Lyapunov equation. If $(\Psi - \Psi_T)$ is positive definite, the system (6) performs better when implemented using finite precision arithmetic. If $Q-Q_T \geq 0$ and $P-TP_T T' \geq 0$, then (7), (8), and the properties of Lyapunov equations imply that $\Psi - \Psi_T \geq 0$ for any choice of C, but $\Psi - \Psi_T \geq 0$ can also obtain when these positivity conditions are not satisfied.

---

†This work was supported in part by the National Institutes of Health under Research Grant Number RR00396 from the Division of Research Resources.

NECESSARY AND SUFFICIENT CONDITIONS FOR QUANTIZATION-ERROR STATISTICS TO BE UNIFORM AND WHITE†

A. B. SRIPAD and D. L. SNYDER
Washington University, St. Louis, Missouri 63130

SUMMARY

A quantizer can be viewed as a non-linear mapping from the domain of continuous-amplitude inputs onto one of the countable number of possible output levels. A statistical model to represent the quantization-error in a quantizer has been developed [1], and the use of such models in analyzing quantization effects in the implementation of digital filters has been studied [2]. The output of a roundoff quantizer x_q can be modelled as an infinite-precision input x and an additive noise ε. Let

$$\varepsilon(i) = x(i) - x_q(i), \quad \text{for all } i \in I.$$

First-Order Statistics.

Widrow [1] showed that under the condition that the input random variable has certain band-limited characteristic function, the additive noise ε is uniformly distributed over the quantization interval Δ; this is frequently referred to as the 'Quantization Theorem'. The bandlimitedness assumption is only a sufficient condition which is not satisfied in many cases. In this paper a weaker sufficient condition which is also necessary for the validity of the uniform noise model is given. The details and proofs are omitted here, but can be found in [3].

THEOREM 1. The probability density of the quantization error is

$$f_\varepsilon(E) = \begin{cases} \frac{1}{\Delta} + \frac{1}{\Delta} \sum_{n \neq 0} \phi_x\left(\frac{2\pi n}{\Delta}\right) \exp\left(-\frac{j2\pi nE}{\Delta}\right), & -\Delta/2 \leq E < \Delta/2 \\ 0, & \text{otherwise} \end{cases}$$

where $\phi_x(u) = E[\exp(iux)]$ is the characteristic function of the input.

THEOREM 2. The characteristic function satisfies $\phi_x(2\pi n/\Delta) = 0$, for all $n \neq 0$, if and only if $f_\varepsilon(E)$ is uniform.

Comments: 1) The quantization error has a zero mean and variance $\Delta^2/12$. 2) The 'Quantization Theorem' is a special case of Theorem 2. 3) Theorem 2 expands the class of input distributions for which uniform noise model can be used. 4) An expression for the error committed in the use of a uniform noise model when Theorem 2 is not satisfied is given from Theorem 1 by $f_\varepsilon(E) - 1/\Delta$.

Second-Order Statistics.

In order to study the second-order statistics of the quantization-error sequence, we make use of the two-dimensional version of the 'Quantization Theorem', which is an extension of the one-dimensional case [1]. We develop a weaker sufficient and necessary condition for a uniform noise model.

THEOREM 3. The joint characteristic function of the input random variables satisifes $\phi_{x_1 x_2}(2\pi\ell/\Delta, 2\pi k/\Delta) = 0$, for $\ell \neq 0$ and $k \neq 0$, if and only if the joint probability density of the quantization-error is uniform.

Comments: 1) The quantization-error sequence is 'white' if Theorem 3 is satisfied. 2) If Theorem 3 is satisfied and, in addition, $\phi_x(2\pi n/\Delta) = 0$ for all $n \neq 0$, it can be further shown that the input and the quantization-error are uncorrelated.

References. [1] B. Widrow, "Stat. Anal. of Ampl. Quant. Sampled-Data Systems," Trans. AIEE, Jan. 1960. [2] A. Oppenheim and R. Schafer, *Dig. Sig. Proc.*, Prentice Hall, 1975. [3] A. Sripad and D. Snyder, "A Nec. and Suff. Cond. for Quant. Errors to be Unif. and White," sub. to IEEE Trans. ASSP.

---

† This work was supported by the National Science Foundation under Grants ENG 74-07800 and ENG 76-11565 and by the National Institutes of Health under Grant RR 00396 from the Division of Research Resources.

DIGITAL OSCILLATORS WITH INHERENT NONLINEARITY

HANOCH UR and GERSHON LESHEM
School of Engineering
Tel Aviv University, Israel

ABSTRACT

The paper considers the digital oscillators in which, similarly to analog ones, a nonlinearity is purposely introduced to achieve amplitude stability. Difference equations similar to the differential equation of Van der Pol are introduced, compared and investigated for amplitude dynamics and stability and harmonic content (distortion).

1. INTRODUCTION

The wide-spread use of digital systems together with the expanding replacement of analog systems by digital ones, has generated the interest in the implementation of digital oscillators. (Namely digital means of generating the series $\cos(n\omega T_s+\psi)$, $n=0, 1, 2 \ldots.$). The digital oscillators begin to span the wide usage and variety of functions and performances of the analog oscillators, and in addition are called upon to fulfill functions such as very high precision, low distortion, etc. Considerable work has already been done in the field of digital oscillators. An application oriented survey [1] includes discussion of quantization effects, their influence on the spectrum and experimental results as well as hardware realization. Mitra and Montgomery [2] investigated the nonlinear effect of accumulation saturation or amplitude stability. Furuno et al. [3] investigated the inner periodicity of digital oscillations as affected by sampling and quantization.

The customary method for implementing a digital oscillator is the solution of the linear harmonic equation. As in the continuous case the amplitude of this solution is entirely dependent on the initial conditions and changes with any disturbance. This phenomenon is overcome in the analog case by the introduction of non-linear elements. The work on nonlinear digital oscillators which appeared in the literature considered only the cases where the nonlinearity was caused by necessity such as quantization or overflow [1,2]. In this paper, we investigate the possibility of the purposeful introduction of nonlinear terms to achieve the same ends of stable amplitude and independence of initial conditions of digital oscillators.

The starting point is the obtaining of a discrete equation which is similar to the Van der Pol one. This is not unique, mainly as one may select, in the creation of the non-linear terms, various values for i in $y(n+i)T$ $i=0, 1, 2,$ and use numerous approximations to the exponential. Proper selection will lead to equations whose solution is less time-consuming on one hand or ones that yield better to analytical treatment on the other (the various equations differ also, of course, in stability, etc.). The steady state is investigated by the harmonic balance method a special formalization of which (proposed by Cypkin [4]) is used to estimate the harmonic content (distortion) as a function of ε , the coefficient of the nonlinear part. Amplitude stability is investigated (and demonstrated) by the small perturbations method.

This paper considers non-linear elements of the instantaneous (memoryless) type. With this type of elements, a given requirement for amplitude stability determines ε, which in turn introduces a given undesirable distortion. This can be overcome in the analog case by utilizing "time averaging" [5]. This idea is followed for digital oscillators in a sequel paper.

2. THE DISCRETE COUNTERPART OF THE VAN DER POL EQUATION

Via "Nonlinear damping" The difference equation for the linear digital Oscillator is:

$$y(n+2) = 2\cos\omega T_s \cdot y(n+1) - y(n); \quad y(o) = a \; ; \quad y(1) = b\cos\omega T_s \triangleq b' \qquad (2.1)$$

The solution of (2.1) is given by:

$$y(n) = a\cos(n\omega T_s) + \qquad (b-a)\cot T_s \cdot \sin(n\;T_s) \qquad (2.2)$$

The amplitude of the solution is very sensitive to a change in i.c. and to any noise introduced in the system. This problem is well known in the continuous cases, and the usual remedy is the introduction of nonlinear characteristics, which ensure independence of the steady state amplitude of the i.c. or disturbances.

The usual analytical approach to these nonlinear phenomena is via the Van der Pol equation [6]:

$$\frac{d^2y(t)}{dt^2} - \varepsilon(1-y^2(t))\frac{dy(t)}{dt} + y(t) = 0 \qquad (2.3)$$

The expression $\varepsilon(1-y^2(t))\dot{y}$ can be viewed as a controlled damping term, so that when the oscillators amplitude is increased, the damping terms changes sign and the oscillations are damped. Thus this system can be considered as a quasilinear one, whose poles move in a narrow band around the $j\omega$ axis. This idea can be used also in the digital systems. We try to build a digital system whose poles in the z plane will lie in a narrow ring around the unit circle. One possibility is to start with a difference equation with a constant damping ([7]) term and change it to a nonlinear one, similar to that of (2.3). In this manner one obtains the following equation

$$y(n+2) = 2\cos\omega T_s \cdot \exp(-\varepsilon(y^2(n+1) - 1))y(n+1) - \exp(-2\varepsilon(y^2(n+1)-1))y(n), \quad \varepsilon > 0 \qquad (2.4)$$

In contrast to the continuous case, the damping appears here in two places, so that it is not clear whether $y^2(n+1)$ or $y^2(n)$ should be used in the second term of the right end side of (2.4). The next paragraph will consider this point.

Via Difference Operators:

All the terms in a differential equation of the type we consider have the same time t as their argument. When replacing derivatives with difference-expressions the time instant is replaced by a number of time samples. Thus for the lower order differentials it becomes necessary to select the exact time argument within the interval. It appears that one effective way is to approximate $\dot{y}$, $\ddot{y}$ in (2.3) by the forward differentials [8] and $y(t)$ by $y(n+1)$; thus:

$$y(t) \to y(n+1) \quad ; \quad \dot{y}(t) \to \Delta y(n) = \frac{1}{T_s}\left(y(n+1) - y(n)\right)$$

$$\ddot{y}(t) \to \Delta^2 y(n) = \frac{1}{T_s^2}\left(y(n+2) - 2y(n+1) + y(n)\right) \tag{2.5}$$

It is interesting to note that if we use the mixed "backward-forward" operator $\Delta\nabla$

$$\Delta\nabla[y(n)] = \Delta\nabla[y(n)] = \frac{1}{T_s^2}\left(y(n+1) - 2y(n) + y(n-1)\right) \tag{2.6}$$

as an approximation for the second order differential, the backward operator ∇ for the first order one and $y(n)$ for $y(t)$ we obtain the same result as in (2.5). Thus if we use (2.5) on (2.3) we obtain:

$$y(n+2) - 2y(n+1)\left[1 + \frac{1}{2}\varepsilon T_s\left(1 - y^2(n+1)\right)\right] + (\omega T_s)^2 y(n+1) + \\ + y(n)\left[1 + \varepsilon T_s\left(1 - y^2(n+1)\right)\right] = 0 \tag{2.7}$$

We should note that (2.7) could be obtained from (2.4) by the approximation

$$\cos\omega T_s \cdot \exp\left(\frac{\varepsilon}{2} T_s\left(1 - y^2(n+1)\right)\right) \approx 1 - \frac{1}{2}(\omega T_s)^2 + \frac{\varepsilon}{2} T_s\left(1 - y^2(n+1)\right) \tag{2.8}$$

Thus if we want (2.4) to correspond to the Van der Pol equation we take $y(n+1)$ and not $y(n)$. Use of $y(n)$ leads as shown later to a different steady state value.

The Multinomial Form:

Equation (2.4) contains "mixed" products of the form: $y(n)y(n+1)$. It is sometimes easier to analyze equations which do not contain such "mixed" products. The "pure" type is called "Multinomial". We thus modify equation (2.4) to the form

$$y(n+2) = 2\cos\omega T_s \cdot \exp\left[\varepsilon\left(1-y^2(n+1)\right)\right]y(n+1) - \exp\left[2\varepsilon\left(1-y^2(n)\right)\right]y(n) \tag{2.9}$$

If ε is assumed to be small ($\varepsilon \ll 1$) equation (2.9) can be rewritten as

$$y(n+2) - (2\cos\omega T_s)(1+\varepsilon)y(n+1) + (1+2\varepsilon)y(n) + 2\varepsilon(\cos\omega T_s)y^3(n+1) - 2\varepsilon y^3(n) \approx 0 \tag{2.10}$$

The last equation can be transformed to state space form (Fig.1) by

$$\underline{X}(n+1) = \underline{A}\underline{X}(n) + \underline{R}\underline{X}^3(n) \; ; \quad X_1(n) = y(n) \quad , \quad X_2(n) = y(n+1)$$

where:

$$A = \begin{pmatrix} 0 & 1 \\ -(1+2\varepsilon) & 2(1+\varepsilon)\cos\omega T_s \end{pmatrix} , \quad R = \begin{pmatrix} 0 & 0 \\ 2\varepsilon & -2\varepsilon\cdot\cos\omega T_s \end{pmatrix} \tag{2.11}$$

3. ANALYSIS OF THE NONLINEAR MODEL

The Steady State Amplitude:

First, we try to evaluate the steady state amplitude of (2.4). By assuming $\varepsilon \ll 1$ we can write approximately:

$$y(n+2) - (2\cos\omega_1 T_s)(1+\varepsilon)y(n+1) + (1+2\varepsilon)y(n) + 2\varepsilon(\cos\omega_1 T_s)y^3(n+1) - \\ - 2\varepsilon y^2(n+1)y(n) = 0 \tag{3.1}$$

Using the "harmonic balance" methods it is assumed that the solution of (3.1) is a sinusoidal function, $y(n) = A\cos(n\omega T_S)$ and that for a zero order approximation A and ω are constants. Substitution of $y(n)$ into (3.1), using the trigonomatic identity $\cos^3 x = 1/4(\cos 3x + 3\cos x)$ and ignoring harmonics of third order, gives after comparing terms of $\sin(n\omega T_s)$ and $\cos(\omega n T_s)$:

$$-\sin 2\omega T_s + 2(1+\varepsilon)\sin\omega T_s \cdot \cos\omega_1 T_s - \frac{3}{2}\varepsilon A^2 \sin\omega T_s \cdot \cos\omega_1 T_s + \varepsilon A^2 \sin 2\omega T_s + + \varepsilon A^2 \sin 2\omega T_2 - \frac{3}{2}\varepsilon A^2 \sin 2\omega T_s = 0 \tag{3.2}$$

$$\cos 2\omega T_s - 2(1+\varepsilon)\cos\omega T_s \cdot \cos\omega_1 T_s + (1+2\varepsilon) + \frac{3}{2}\varepsilon A^2 \cos\omega T_s \cdot \cos\omega_1 T_s - -\varepsilon A^2(1-\cos 2\omega T_s) - \frac{3}{2}\varepsilon A^2 \cos 2\omega T_s = 0 \tag{3.3}$$

Thus, $A = 2 \quad , \quad \omega = \omega_1$ (3.4)

Note that this amplitude is identical to that of the solution of the Van der Pol equation in the continuous case. In a similar manner the steady state value for the multinomial form equation (2.10) is found to be :

$$A_m = \sqrt{\frac{4}{3}} = 1.157 \quad . \tag{3.5}$$

Envelope Stability - Small Perturbation Method

In order to investigate the dynamics of the envelope we introduce the familiar envelope-phase representation (assuming zero i.c.)

$$y(n) = R(n)\cos[\theta(n)] = R(n)\cos\omega' T_s \quad ; \quad (\omega' \neq \omega) \ , \ \theta(o) = 0 \tag{3.6}$$

We substitute equation (3.6) into equation (2.10) and comparing coefficients of $\sin(n\omega T_S)$ and $\cos(n\omega T_S)$, we get two equations. Substituting one into the other:

$$R(n+2) - (1+2\varepsilon)R(n) + \frac{3}{2}\varepsilon R^3(n) = 0 \qquad (\varepsilon << 1) \tag{3.7}$$

It is easy to see that the substitution $R_{SS} = \pm\sqrt{\frac{4}{3}}$ in (3.7) gives zero for every ε, (equation(3.7), is true however only for $\varepsilon << 1$). If we substitute $\varepsilon = 0$ in equation (3.7) (the linear case) we obtain $R(n+2) = R(n) \rightarrow \lambda_{1,2} = \pm 1$ (two roots on the unit circle), so that the envelope is not asymptotically stable ("neutral limit cycle"). The stability of the envelope for $\varepsilon > 0$ can be investigated around the equilibrium point $R_{SS} = \sqrt{\frac{4}{3}}$ by using a small perturbation around it. (3.7) is first changed into state space form and then the Jacobian - J - is evaluated around R_{SS}, from which one obtains the characteristic equation

$$|J - \lambda I| = \begin{vmatrix} -\lambda & 1-4\varepsilon \\ 1 & -\lambda \end{vmatrix} = 0 \qquad \lambda_{1,2} = \pm\sqrt{1-4\varepsilon} \tag{3.8}$$

which can be described by root locus of Fig. 2 (ε parameter), which shows that for small increasing ε the envelope becomes more stable, and unstable again for large ε . We have to remember that the analysis is valid only for $\varepsilon << 1$, and comparison with computer simulation shows that it is so for $\varepsilon < 0.1$. This simulation also shows that for large ε the solution is no more periodic, in contrast to the analog case. The small perturbation method was applied here to the envelope of (2.10). It may also be applied to that of (3.1) with similar results [9].

Harmonic Analysis

In this section we will investigate the steady state via the Cypkin method [4]. The method (Appendix A) is based on representing the steady state solution of the multinomial nonlinear difference equation (in the form of eq. A1) of the type of (3.9) by a Fourier Series.

$$y(n+2) + \alpha y(n+1) + \beta y(n) + \gamma y^3(n+1) + \delta y^3(n) = X(n) = 0 \tag{3.9}$$

The nonlinear equation can be represented by the feedback system of Fig. 3. The linear part of the equation being represented by H_L, (inner low pass filter) and because of the special form of (A1) or (3.9) only one nonlinear block in Fig. 3 represents all the nonlinear terms. Manipulation of (A1) yields in the general case

$$H_L = \frac{\Sigma \beta_i z^{-i}}{\Sigma \alpha_i z^{-i}} \tag{3.10}$$

In our case, using the notations of (3.9) and Appendix A:

$$H_L(z) = \frac{2\varepsilon(1 - \cos\omega T_S \cdot z)}{(1+2\varepsilon) - 2(1+\varepsilon)\cos\omega T_S)z + z^2} \tag{3.11}$$

$$Q(jk\omega) = (1+2\varepsilon) - 2(1+\varepsilon)\cos\omega T_s \cdot \exp(jK\omega) + \exp(2jK\omega)$$

$$P(jK\omega) = 2\varepsilon(1 - \cos\omega T_S \cdot \exp(jK\omega)\ ) \tag{3.12}$$

Using the Cypkin method we can now find the first and the third harmonics (complex amplitudes), a'_1 , a'_3 ; for $\varepsilon << 1$. We assume arbitrarily that $\phi_1 = 0$, so: $a'_1 = a_1$, and first take

$$y(n) \approx a_1 \cos(n\omega T_s + \phi_1) \tag{3.13}$$

$$\left.\frac{P(jk\omega)}{Q(jk\omega)}\right|_{k=1} \triangleq \frac{Q_1}{P_1} = \frac{2}{Na_1} \sum_{i=o}^{N-1} \hat{\phi}(a_1 \cos(\frac{\pi}{N} i) \exp(-j\omega i) = \tag{3.14}$$

$$= \frac{2a^2{}_1}{N} \left[\sum_{i=o}^{N-1} \cos^4(\frac{\pi i}{N}) - j \sum_{i=o}^{N-1} \cos^3(\frac{\pi i}{N}) \sin(\frac{\pi}{N} i) \right]$$

The first sum is 3/8 N, while the second is zero, thus

$$\frac{Q_1}{P_1} = \frac{3}{4} a_1^2 \tag{3.15}$$

But from (3.12) : $Q_1/P_1 = 1$, so $a_1 = \sqrt{4/3} = R_{ss}$, which agrees with the result previously obtained. To find a'_3 we assume first a solution of the form :

$$y(n) \approx a_1 \cos(\frac{\pi}{N} n) + a_3 \cos(3 \frac{\pi}{N} n + \phi_3); \quad \frac{\pi}{N} = \omega T_s \tag{3.16}$$

(A.4) and (A.5) $a'_3 = a_3 \exp(j\omega_3) = \frac{2P_3}{NQ_3} \sum_{i=o}^{N-1} y^3(i) \exp(-j3\omega i)$ (3.17)

We define $S(a_3, \phi_3, N) = \frac{a_3 Q_3}{P_3} = \frac{2\exp(-j\phi_3)}{N} \sum_{i=o}^{N-1} y^3(i) \exp(-j3\omega i)$ (3.18)

We can expand $S(a_3, \phi_3, N)$ in series about $a_3 = o$ (for $\varepsilon << 1, |a'_3/a'_1| << 1$) and define :

$$S|_{a_3=o} \triangleq S_{o3}\ ; \quad P_3 \triangleq \varepsilon P_{o3}\ ; \quad P_{o3} \triangleq 2(1-\cos\omega T_s \cdot \exp(j3\omega T_s)\ ); \quad Q_3 \triangleq P_3 + Q_{o3} \tag{3.19}$$

$$\frac{S(a_3,\phi_3,N)}{a_3} = \frac{S_{o3}}{a_3} + \frac{\partial s_{o3}}{\partial a_3} + \;.\;.\;.\;.\; = \frac{Q_3}{P_3} = 1 + \frac{Q_{o3}}{P_{o3}}\,\frac{1}{\varepsilon} \qquad (3.20)$$

under assumptions of smallness of ε, a_3 :

$$\frac{1}{\varepsilon}\,\frac{Q_{o3}}{P_{o3}} \approx \frac{S_{o3}}{a_3} \;\therefore\; a_3 = \varepsilon\,\frac{P_{o3}}{Q_{o3}} \cdot S_{o3} \qquad (3.21)$$

Taking the expression for S_{o3} from (3.18) performing the sums one obtains

$$S_{o3} = \frac{2\exp(-j\,\phi_3)}{N} \sum_{i=o}^{N-1} a_1^3 \cos^3\left(\frac{\pi}{N}\,i\right)\exp(-j3\omega i) = \frac{a_1^{\,3}}{4}\exp(-j\phi_3) \qquad (3.22)$$

Thus

$$a_3' = \varepsilon\frac{a_1^{\,3}}{4}\;\frac{P_{o3}}{Q_{o3}} = \varepsilon\,\frac{P_{o3}}{Q_{o3}}\,\frac{1}{3}\sqrt{\frac{4}{3}} \qquad (3.23)$$

We can now define the distortion coefficient ρ and obtain its value:

$$\rho \overset{\Delta}{=} \left|\frac{a_3'}{a_1'}\right| = \varepsilon\,\frac{1}{3}\left|\frac{P_{o3}}{Q_{o3}}\right| \qquad (3.24)$$

Numerical Example

Let $N = 20$ (thus $\omega T_s = 9°$, and $\cos\omega T_s = 0.9877$) from (3.19):

$$P_{o3} = 0.2399 - j0.8968 = 0.9283\,\underline{/-75°}\,,\; Q_{o3} = 0.1932\,\underline{/26.98°}\,,\; a_3 = 1.84\varepsilon\,,$$

$$\phi_3 = -101.98°\,,\; a_1 = 1.154 = \sqrt{\frac{4}{3}}\,,\; \phi_1 = 0°$$

Thus in the present example: $\rho = 1.6\varepsilon$ (1.6% for .01)

4. SUMMARY, COMPUTER RESULTS AND CONCLUSIONS

This paper considers the effects, both desireable and undesireable, of introduction of nonlinearities in digital oscillators. The main reason for introduction is amplitude regulation; in Fig. 4a we show the output of a linear oscillator; with initial value of amplitude 1.0 and an impulse disturbance of 5.0. The amplitude jumps to about 30. Fig. 4b shows (note the change in scale) the quick recovery, under same conditions, when we have a nonlinear term ($\varepsilon = 0.01$)(Eq. 2.9). The phase-plane graph of Fig. 5 shows the convergence of the amplitude to the steady state value, from arbitrary initial conditions. We pay, of course, a price - distortion, Fig. 6, where a larger value of ε (0.1) was taken for demonstration purposes. The value of the distortion - the harmonic content - was estimated analytically by harmonic balance (Cypkin Method). As mentioned earlier, it may not be possible to obtain simultaneously good amplitude regulation and low distortion. This may be achieved as in the analog case [5] by extra processing, which increases the filtering effect, thus assuring low distortion even for large ε. The method, "time averaging" is applied in a following paper. The envelope stability was investigated by small perturbations.

APPENDIX A: CYPKIN METHOD [7]

The basic equation (similar to (3.9)) is

$$\sum_{i=o}^{m} \alpha_i y(n+1) = \sum_{i=o}^{m'} \beta_i \hat{\phi}\,\{y(n+1)\}\;;\quad m > m';\quad T/T_s > 3 \qquad (A.1)$$

where $\hat{\phi}$ is nonlinear anti-symmetric function.

If $y(n)$ is a periodic function, $\hat{\phi}(y)$ is also periodic in n, so it can be described by a Fourier series, thus

$$y(n) = \frac{1}{2} \sum_{K=-N_1}^{N_1} a'_K \exp(jK\omega n), \quad N_1 = \begin{cases} N-1, & N \text{ even}, \\ N, & N \text{ odd}. \end{cases} \quad ; \quad \omega T_s = 2\pi \frac{T_s}{T} \triangleq \frac{\pi}{N} \qquad (A.2)$$

So N relates to half a period, k takes only odd values, and a'_k are the complex amplitudes $a'_k = a_k \exp(j\phi_k)$. Thus we can write:

$$\hat{\phi}[y(n)] = \hat{\phi}\left[\frac{1}{2} \sum_{K=-N_1}^{N_1} a'_k \exp(jk\omega n) \right] = \frac{1}{2} \sum_{K=-N_1}^{N_1} b'_k \exp(jk\omega n) \qquad (A.3)$$

b'_k can be found by harmonic Bessel analysis ([4]):

$$b'_k = b_k \exp(j\phi_k) = \frac{2}{N} \sum_{i=0}^{N-1} \hat{\phi}\,(y(i))\exp(-jk\omega i) \qquad (A.4)$$

We now substitute $y(n)$ and $\hat{\phi}(y)$ in (2.10). Comparing coefficients of $\exp(j\omega n)$:

$$a'_k Q(jk\omega) = b'_k P(jk\omega), \; k = 1,3,\ldots,; \quad Q(jk\omega) = \sum_{i=0}^{m} \alpha_i \exp(jk\omega i)$$

$$P(jk\omega) = \sum_{i=0}^{m'} \beta_i \exp(jk\omega i) \qquad (A.5)$$

The general (steady state) solution of (A.1) in the interval: $0 \le n \le N-1$ (half period) is obtained by substituting (A.4),(A.5) in (A.2):

$$y(n) = \frac{1}{N} \sum_{i=0}^{N-1} A_{n-i}(N)\hat{\phi}[y(i)] \qquad (A.6)$$

where:

$$A_{n-i}(N) = \sum_{K=-N+2}^{N-2} \frac{P(jk\omega)}{Q(jk\omega)} \exp[jk\omega(n-i)] + \frac{P(j\pi)}{Q(j\pi)} \exp[j\pi(n-i)], \; (N \text{ even}) \qquad (A.7)$$

It can be written, using H_L, as:

$$A_{n-i}(N) = \sum_{K=-N+2}^{N} H_L(jk\omega)\exp[jk\omega(n-i)] \qquad (A.8)$$

So, in order to obtain the periodic solution of the nonlinear equation (2.10), we must solve a set of nonlinear algebraic equations.

REFERENCES

[1] J. Todtli, T. Pfund, "Digitler Oscillator," Agen-Mitteilungen, Institut für Technische Physik der Eidgenossischen Technischen Hochschule, Zurich, Switzerland, 1970.

[2] D. Mitra, H.D. Montgomery, "On a Digital Oscillator," Bell Laboratories, Murray Hill, New Jersey, 1972.

[3] K. Furuno, S.K. Mitra, K. Hirano, Y. Ito, "Design of Digital Sinusoidal Oscillators with absolute periodicity," IEEE, Proc. 1975, ISCAS.

[4] P. Vidal, "Non-linear Sampled-Data Systems," Gordon & Breach, 1969.

[5] H. Ur, T. Lissack, "Time Averaged Non-linear Systems, Steady-state Analysis," Int. J. Control, Vol. 17, No. 6, 1143-1152, 1973.

[6] N. Kryloff, N. Bogoluboff, "Introduction to Non-linear Mechanics," Princeton Univ. Press, 1947.

[7] Gold and Rader, "Digital Processing of Signals," 1969.

[8] M. Cuendo, A. Durling, "A Discrete Time Approach for System Analysis," Academic Press, 1969.

[9] G. Leshem, "Non-linear Digital Oscillator with and without Time Averaging, M.Sc. Thesis, Tel-Aviv University, 1976.

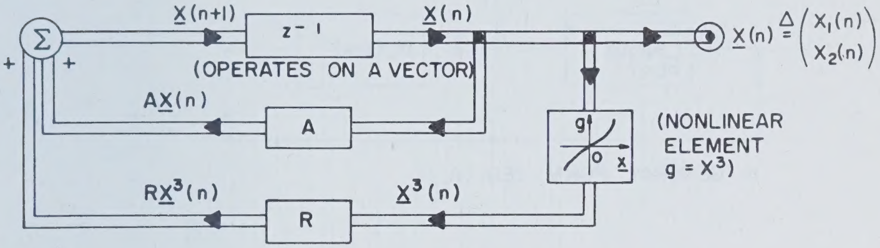

Σ
+
+
X(n+1)
z⁻¹
X(n)
(OPERATES ON A VECTOR)
X(n) ≜ (X₁(n), X₂(n))
AX(n)
A
g
0
x
(NONLINEAR ELEMENT g = X³)
RX³(n)
R
X³(n)

Figure 1 : A possible implementation of equation (2.13).

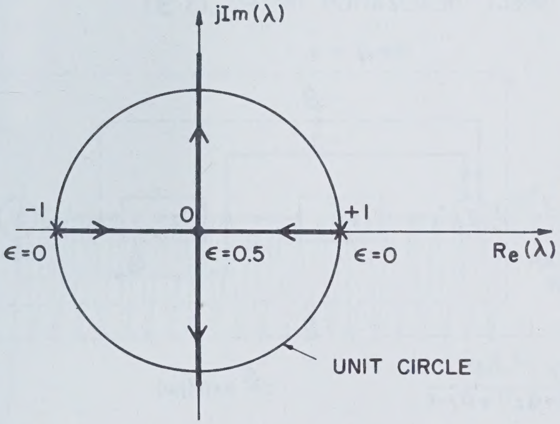
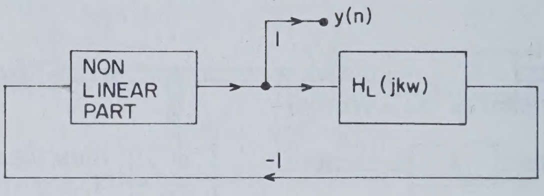
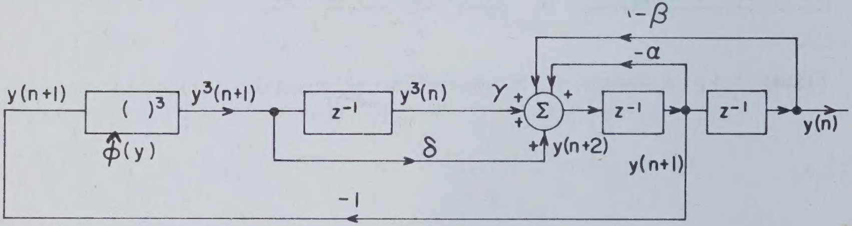

jIm(λ)
-1
0
+1
ε=0
ε=0.5
ε=0
Re(λ)
UNIT CIRCLE

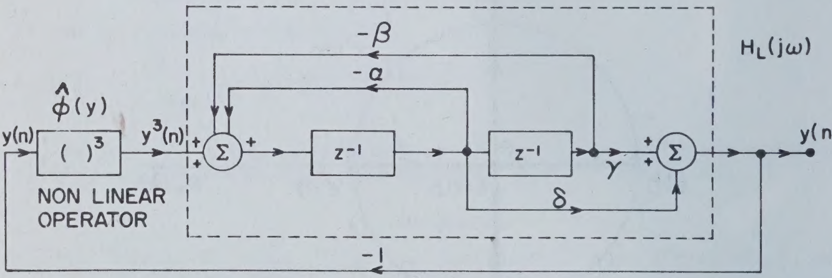
Figure 2 : Root locus of the envelope for the parameter ε equation (3.7).


y(n)
1
NON
LINEAR
PART
H_L(jkw)
-1

a. GENERAL FORM (EQ. (A·1)


-β
-α
y(n+1)
( )^3
$\hat{\phi}(y)$
y^3(n+1)
z^-1
y^3(n)
γ
+
+
+
Σ
+
y(n+2)
δ
z^-1
z^-1
y(n)
y(n+1)
-1

b. DIRECT REALIZATION OF EQ. (3·9)


H_L(jω)
-β
-α
$\hat{\phi}(y)$
y(n)
( )^3
y^3(n)
+
+
Σ
+
z^-1
z^-1
γ
+
+
Σ
δ
y(n)
NON LINEAR
OPERATOR
-1

$$H_L(j\omega) = \frac{-\gamma z^{-2} - \delta z^{-1}}{1 + \alpha z^{-1} + \beta z^{-2}} \qquad z \overset{\Delta}{=} \exp(j\omega)$$

c. CANONIC REALIZATION OF EQ. (3·9)

Figure 3 : Feedback realizations for harmonic analysis.

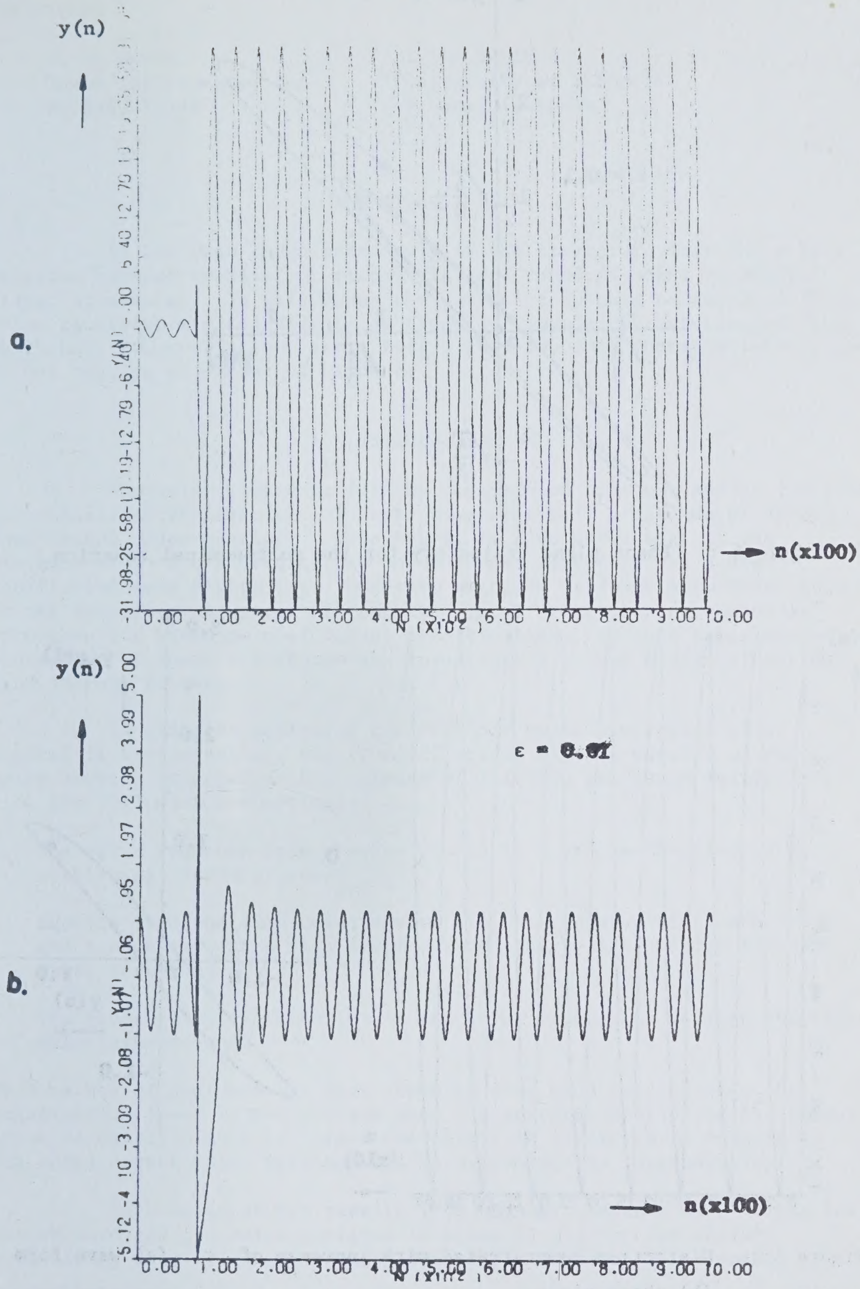


Figure 4. The effect of a disturbance as the output of a digital oscillator, initial amplitude 1, impulse disturbance 5. (a) linear oscillator (b) non-linear ($\varepsilon = 0.01$) oscillator.

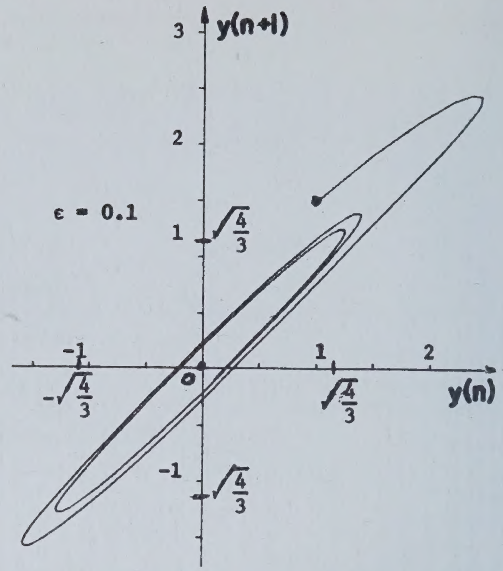


Figure 5 : Phase plane trajectory for the multinominal equation.

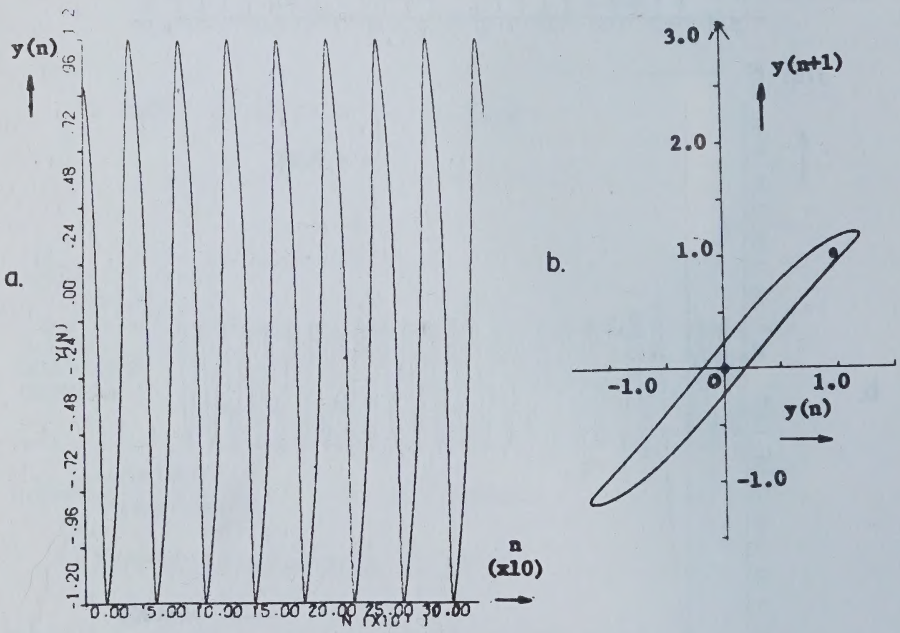


Figure 6 : Distortion demonstrated with increase of ε. (a) wave form (b) phase plane.

SOME LOWPASS NARROW-BAND LOW-NOISE FIXED-POINT RECURSIVE DIGITAL FILTER STRUCTURES[†]

D. J. MAYER
Texas Instruments, Inc.
Dallas, Texas

T. N. TRICK
University of Illinois
Urbana, Illinois

ABSTRACT

It has been shown previously in the literature that there is a relation between coefficient sensitivity and roundoff noise in digital filter structures. In this paper the design of lowpass narrow-band fixed-point recursive digital filters is studied. Several structures are found which have low coefficient sensitivity, good roundoff characteristics, and do not require an excess of hardware.

1. INTRODUCTION

A popular structure for the realization of the transfer function of a linear shift-invariant discrete time system is a cascade of direct form second order sections. However, it is also known that in the vicinity of $z = 1$ the direct form second order structure has a high coefficient sensitivity [1]. However, multiple feedback structures such as the leap-frog structure [2] and the second order coupled resonator structure [1] have low coefficient sensitivities. In this paper the noise properties of these structures are investigated in the design of narrow-band lowpass filters.

In order to determine the roundoff noise properties of a digital filter structure, the roundoff errors will be modeled as random noise sources injected at the outputs of infinite precision multipliers with the following assumptions:

1. The error sequence from a noise source is a sample function of a stationary random process.
2. Samples from the same noise source are uncorrelated with each other and are uncorrelated with samples from other noise sources and the input sequence.
3. The probability distribution of the error process is uniform over the quantization range.

This method of analysis has been shown to give good results when the quantization level is not low and when the spectral contant of the filter input is fairly high [3]. Since the filter is linear shift invariant, the total output noise variance can be determined by superposition.

Each noise source results from the quantization of a multiplier output word and its noise variance is given by (1) for the uniform probability density function where d is the quantization increment.

[†] This work was supported by the Joint Services Electronics Program (U.S. Army, U.S. Navy, U.S. Air Force) under contract DAAB-07-72-C-0259.

$$\sigma_o^2 = d^2/12 \qquad (1)$$

Rounding will be assumed so that the mean is zero.

The variance of the noise at the filter output due to noise source i is given by

$$\sigma_i^2 = \sigma_o^2 \sum_{k=-\infty}^{\infty} |h_{ni}(k)|^2 \qquad (2)$$

where $h_{ni}(k)$ is the impulse response from the point where the noise is injected to the filter output. The output noise variance can be represented as a ratio of σ_i^2 to σ_o^2 to eliminate the dependence on the quantization increment d. Equations (3) and (4) are equivalent to (2) and are generally of more use in noise calculations [1].

$$\frac{\sigma_i^2}{\sigma_o^2} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H_{ni}(e^{j\omega})|^2 d\omega \qquad (3)$$

$$\frac{\sigma_i^2}{\sigma_o^2} = \frac{1}{2\pi j}\int H_{ni}(z)H_{ni}(z^{-1})z^{-1}dz \qquad (4)$$

$H_{ni}(z)$ is the transfer function in the z-transform domain from the noise source to the output and $H_{ni}(e^{j\omega})$ is the corresponding Fourier transform transfer function. The total noise variance at the filter output is found using superposition and is given by

$$\frac{\sigma_n^2}{\sigma_o^2} = \sum_{i=1}^{M} \frac{\sigma_i^2}{\sigma_o^2} \qquad (5)$$

where M is the number of noise sources.

Before the noise variance of a filter can be calculated, overflow constraints must be imposed and met at each summation point. This is accomplished by scaling appropriate multipliers in the filter. The theoretically calculated variance of an improperly scaled filter can be radically different from that of the same filter properly scaled. Various methods for scaling filters exist and each yields slightly different scaling multipliers [4]. In this paper scaling is accomplished by satisfying the constraint

$$\max_{-\pi<\omega<\pi} |G_{i1}(e^{j\omega})| = 1 \qquad (6)$$

at each node i where a summation occurs. $G_{i1}(e^{j\omega})$ is the transfer function from the input to the summation node i. This is the L_∞-norm discussed by Jackson [4]. Filters scaled with norms other than that of (6) will have slightly different noise properties, however general trends in the noise performance of a structure should be about the same. Two's complement arithmetic is assumed so that overflows may occur internally at a summation node with more than two inputs while the net sum still meets the requirement of (6).

2. THE COUPLED FORM

The coupled form realization of a complex pole pair as proposed by Gold and Rader [5] is illustrated in Figure 1a. The transfer function for this structure is

$$H(z) = \frac{\alpha_o \alpha_1 z^{-2}}{1 - (\beta_1 + \beta_2) z^{-1} + (\beta_1 \beta_2 - \alpha_1 t) z^{-2}} . \tag{7}$$

For comparison the direct form second order structure is shown in Figure 1b. The transfer function for this structure is

$$H(z) = \frac{\alpha z^{-2}}{1 - 2r \cos\theta \, z^{-1} + r^2 z^{-2}} . \tag{8}$$

Thus these two structures realize the same pole pair if

$$\beta_1 + \beta_2 = 2r \cos\theta$$

and

$$\beta_1 \beta_2 - \alpha_1 t = r^2 .$$

The coefficients α_o, α_1, and α are scaling coefficients and are set by the scaling required to avoid overflow at nodes n_1 and n_2 in Figure 1. Discounting the scaling multipliers α, α_o, and α_1, the coupled form has an extra multiplier and hence an extra degree of freedom. This extra degree of freedom will be used to minimize the noise in the coupled resonator structure.

A noise analysis of these structures yields

$$\left(\frac{\sigma_n^2}{\sigma_o^2}\right)_{\text{coupled}} = \left[\alpha_1^2 + \frac{2}{3}\left(1 + \beta_1^2 - \frac{4r\cos\theta}{1+r^2}\beta_1\right)\right] N_d \tag{9}$$

and

$$\left(\frac{\sigma_n^2}{\sigma_o^2}\right)_{\text{direct}} = N_d \tag{10}$$

where

$$N_d = \frac{3(1+r^2)}{(1-r^2)(1+r^4-2r^2\cos 2\theta)} . \tag{11}$$

Note that N_d is dependent on the pole locations. Hence in order to minimize the roundoff noise of the coupled form β_1 was chosen to minimize the coefficient of N_d in (9) while α_o and α_1 were chosen such that

$$\max_{-\pi < \omega < \pi} |H_1(e^{j\omega})| = \max_{-\pi < \omega < \pi} \left| \frac{\alpha_o(1 - \beta_2 e^{-j\omega})}{1 - 2r\cos\theta e^{-j\omega} + r^2 e^{-2j\omega}} \right| = 1 \tag{12}$$

and

$$\max_{-\pi < \omega < \pi} |H_2(e^{j\omega})| = \max_{-\pi < \omega < \pi} \left| \frac{\alpha_o \alpha_1 e^{-j\omega}}{1 - 2r\cos\theta e^{-j\omega} + r^2 e^{-2j\omega}} \right| = 1. \tag{13}$$

The following computer algorithm was used to minimize the noise in the coupled resonator structure.

1. Select a value for β_1 and solve $\beta_2 = 2r\cos\theta - \beta_1$.
2. Determine α_o from (12).
3. Determine α_1 from (13).
4. Calculate $M = \alpha_1^2 + \frac{2}{3}(1 + \beta_1^2 - \frac{4r\cos\theta}{1+r^2}\beta_1)$.
5. Repeat steps 1-4 until a minimum of M is found.

For comparison rather than select β_1 to minimize the roundoff noise, we also selected $\beta_1 = 1$ in order to reduce the number of multipliers. The roundoff noise of these three structures is shown in Figure 2 as a function of θ with $r = .99$ (high Q pole). Coupled form A is the structure with $\beta_1 = 1$ and coupled form B is the structure in which β_1 was chosen to minimize the noise. The remarkable observation is that coupled forms A and B have about the same noise output. Only for very small values of θ does the optimized structure offer slightly superior performance. For larger values of θ coupled form A is slightly better since it has one less multiplier. Finally, note that as expected the noise performance of the cascade structure is very poor for low values of θ, because its coefficient sensitivity is also poor in this region.

3. A MULTIPLE FEEDBACK STRUCTURE

Another structure which has low coefficient sensitivity with proper design is the leap-frog structure shown in Figure 3. For $N = 3$ the transfer function of the structure is

$$H(z) = \frac{T_1 T_2 T_3}{1 - t_{12}T_1T_2 - t_{23}T_2T_3} . \tag{14}$$

Suppose that the parameter p_2 is contained in T_2, then

$$S_{p_2}^{H} \triangleq \frac{p_2}{H}\frac{\partial H}{\partial p_2} = \frac{1}{1 - t_{12}T_1T_2 - t_{23}T_2T_3} S_{p_2}^{T_2} . \tag{15}$$

Let $T_i = N_i/D_i$, then (15) becomes

$$S_{p_2}^{H} = \frac{D_1D_2D_3}{D_1D_2D_3 - t_{12}N_1N_2D_3 - t_{23}N_2N_3D_1} \cdot S_{p_2}^{T_2} . \tag{16}$$

An examination of (16) reveals that the passband sensitivity can be reduced by placing the open-loop poles D_1, D_2, and D_3 in the passband and on the unit circle of the z plane. We have no control over the denominator of (16), since it is determined by the pole requirement of the transfer function (14) of the filter.

From this observation Long [6] designed some low sensitivity narrow-band lowpass filters using a cascade of first-order sections in the leap-frog structure of Figure 3. Since there are extra degrees of freedom (more feedback multipliers than denominator coefficients) Long chose as many open loop poles as possible to lie at $z = 1$. This not only resulted in zeros of the sensitivity function at $z = 1$, but also reduced the number of multipliers since a pole at $z = 1$ means that $\beta_i = 1$ in the first order transfer function $T_i(z)$. Long's structure for a fifth order Chebyshev lowpass filter is shown in Figure 4. Note that α_{oi} must be set to zero in every other section in order to avoid delay free loops. However, whenever $\alpha_{oi} = 0$ it is impossible to realize a zero at $z = -1$. These zeros must be added at the end of the filter.

In order to include the zeros in the $T_i(z)$ sections and avoid delay free loops the structure in Figure 5 was proposed [7]. Not only did this structure correct the delay-free loop problem, but it was also found that the multiplier coefficients are identical to Long's structure. This is observed from Figure 6. Figure 6a is a block diagram realization of Long's structure, and Figure 6b is a block diagram of the modified structure. From flow graph techniques we note that both have the same input-output relationship provided that

$$N_i = N_{i+1}, \quad \text{for all even } i.$$

This is no problem in the Chebyshev lowpass filter since all the zeros lie at $z = -1$.

4. COMPARISON OF STRUCTURES

In order to compare the noise performance of the coupled form and the leap-frog form with the cascade form two narrow-band Chebyshev lowpass filters were designed. Both are fifth order filters with 0.5 dB passband ripple. The first is designated LP8 since it has a cutoff frequency of π/8 radians where π radians denotes the folding frequency. The second filter is designated LP80 since its cutoff frequency is π/80 radians. Note that the bandwidth of the second filter is 1/10 of the first filter. In all filters the sections were ordered according to the algorithms proposed by Jackson [4].

The results of the noise analysis are illustrated in Table 1. The LP8 filter is not a very narrow-band filter, but still the coupled form realization is approximately 8 dB better than the cascade and the multiple feedback structure was 10.4 dB better. As expected in the very narrow-band filter (LP80) the improvement is even more dramatic. The coupled form realization was approximately 26 dB better than the cascade form. A multiple feedback filter was designed from the LP8 design by applying the lowpass transformation

$$z^{-1} \rightarrow \frac{z^{-1} - a}{1 - az - 1}$$

where

$$a = \frac{\sin(\frac{\theta_c - \omega_c}{2})}{\sin(\frac{\theta_c + \omega_c}{2})} \quad \text{and}$$

θ_c is the cutoff of the prototype filter (LP8) and ω_c is the cutoff of the desired structure (LP80). This substitution was applied directly to the LP8 structure but resulted in a structure with delay free loops and twice as many multipliers. Even though the resulting structure was neither practical nor realizable its noise performance was calculated for the sake of comparison. Note that even with the large increase in multipliers the structure still has a noise performance nearly equal to that of the coupled form and far superior to that of the cascade form. One would expect that with proper design one could achieve a noise performance somewhat better than the coupled form.

Filter	Cascade	Coupled Form A	Coupled Form B	MFB
LP8	26.9	19.2	19.3	16.5
LP80	49.1	22.7	23.0	23.9

Table 1. Normalized Noise Variance σ_n^2/σ_o^2 in dB.

REFERENCES

1. A. V. Oppenheim and R. W. Schafer, Digital Signal Processing, Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

2. G. Szentirmai, "Synthesis of Multiple-Feedback Active Filters," Bell System Technical Journal, Vol. 52, pp. 527-555, April, 1973.

3. L. B. Jackson, "Roundoff Noise Analysis for Fixed-Point Digital Filters Realized in Cascade and Parallel Form," IEEE Trans. on Audio and Electroacoustics, Vol. AU-18, pp. 107-122, June, 1970.

4. L. B. Jackson, "On the Interaction of Roundoff Noise and Dynamic Range in Digital Filters," Bell System Technical Journal, Vol. 49, pp. 159-184, February, 1970.

5. B. Gold and C. M. Rader, Digital Signal Processing of Signals, McGraw-Hill, New York, 1969.

6. J. L. Long and T. N. Trick, "Sensitivity and Noise Comparison of Some Fixed-Point Recursive Digital Filter Structures," Proc. of the IEEE International Symposium on Circuits and Systems, pp. 56-60, 1975.

7. D. J. Mayer, "Roundoff Noise in Fixed-Point Recursive Digital Filters," M.S. thesis, University of Illinois, Urbana, Illinois, October, 1976.

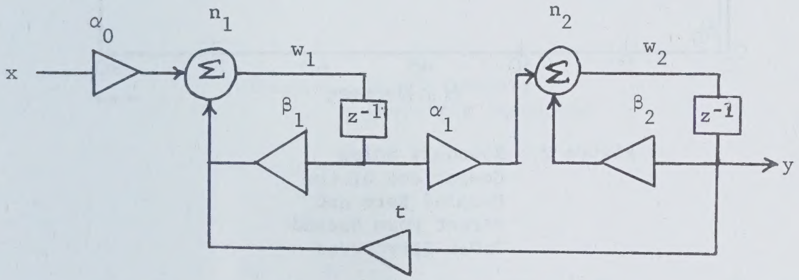

α_0
n_1
n_2
x
Σ
w_1
Σ
w_2
β_1
z^{-1}
α_1
β_2
z^{-1}
y
t

Figure 1a. Coupled Form

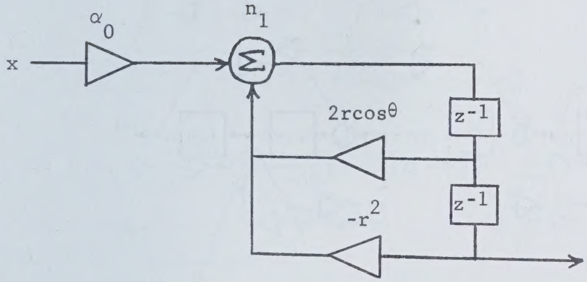

α_0
n_1
x
Σ
$2r\cos\theta$
z^{-1}
z^{-1}
$-r^2$

Figure 1b. Direct Form

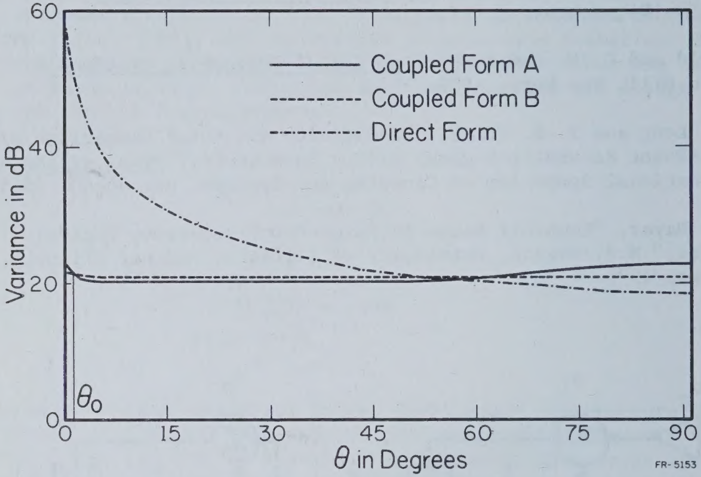


Figure 2. Roundoff Noise Comparison of the Coupled Form and Direct Form Second Order Structures

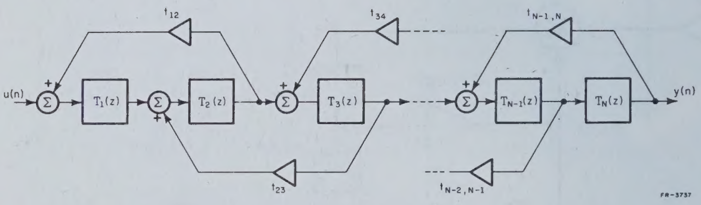
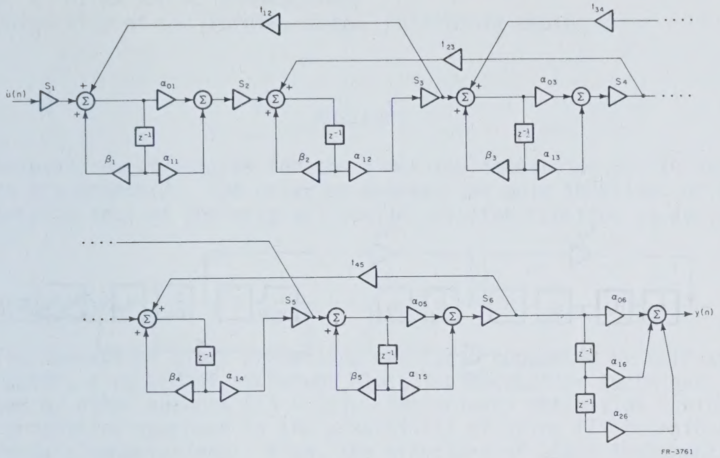
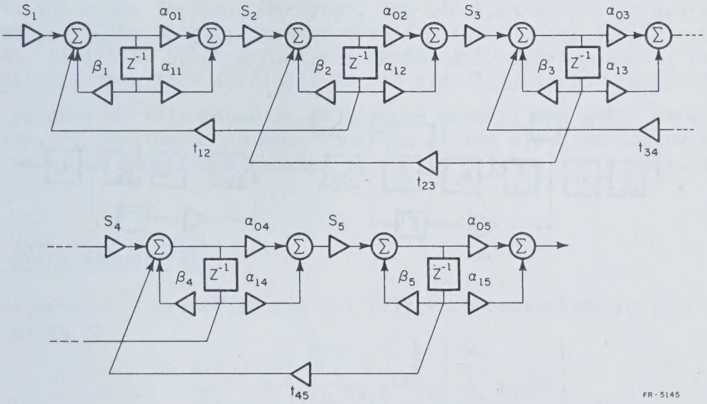

t12
t34
tN-1,N
u(n)
+
Σ
T1(z)
Σ
+
T2(z)
+
Σ
T3(z)
+
Σ
TN-1(z)
TN(z)
y(n)
t23
tN-2,N-1
FR-3737

Figure 3. Leap-Frog Structure



t_{12}
t_{34}
t_{23}
u(n)
S_1
α_{01}
S_2
S_3
α_{03}
S_4
z^{-1}
β_1
α_{11}
β_2
α_{12}
β_3
α_{13}
t_{45}
S_5
α_{05}
S_6
α_{06}
y(n)
β_4
α_{14}
β_5
α_{15}
α_{16}
α_{26}
FR-3761

Figure 4. Long's Lowpass
Leap-Frog Structure



S_1
α_{01}
S_2
α_{02}
S_3
α_{03}
β_1
Z^{-1}
α_{11}
β_2
α_{12}
β_3
α_{13}
t_{12}
t_{34}
t_{23}
S_4
α_{04}
S_5
α_{05}
β_4
α_{14}
β_5
α_{15}
t_{45}
FR-5145

Figure 5. Modified Lowpass
Leap-Frog Structure

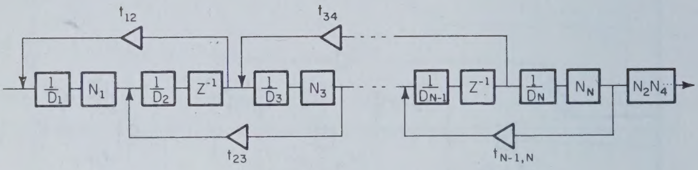

t12
t34
1/D1
N1
1/D2
Z-1
1/D3
N3
1/DN-1
Z-1
1/DN
NN
N2N4...
t23
tN-1, N

Figure 6a. Block Diagram of Long's Structure

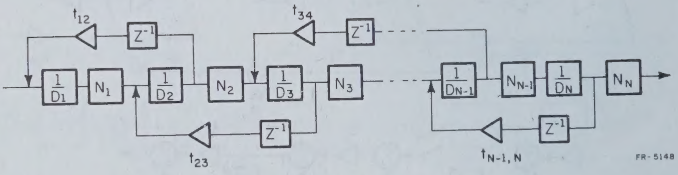

t12
Z-1
t34
Z-1
1/D1
N1
1/D2
N2
1/D3
N3
1/DN-1
NN-1
1/DN
NN
Z-1
t23
Z-1
tN-1, N
FR-5148

Figure 6b. Block Diagram of the Modified Structure

NEW RESULTS ON THE BLOCK IMPLEMENTATION OF RECURSIVE DIGITAL FILTERS

S. K. MITRA AND R. GNANASEKARAN
University of California, Davis, California 95616

ABSTRACT

Several new structures for the block implementation of IIR digital filters are proposed. The relation between the pole locations of the block structure to that of the original scalar transfer function is derived.

I. Introduction

The concept of block processing was first suggested by Gold and Jordan [1]. Later, a number of different block implementation techniques have been proposed by other authors [2] - [6]. The primary motivation behind the block processing approach is the possibility of using FFT techniques for intermediate computations. Also, the structure of block implementation lends itself for computation by the array processors.

An unified and elegant approach for block processing was developed by Burrus [2], [3] by making use of the matrix representation of the convolution. This approach to block implementation also has led to the representation of periodically time-varying digital filters [7] by a time-invariant digital filter. It also finds application in the FFT implementation of multirate digital filters [8]. An attractive feature of the block implementation method proposed by Burrus is that stability of block structure can be easily tested by examining the eigenvalues of a particular L X L matrix where L is the block length. Moreover, the block structure is guaranteed to be stable, if the original scalar transfer function is stable. On the other hand, in all of block structures proposed by other authors, the stability analysis is either difficult or the stability cannot be guaranteed.

The purpose of this paper is to advance several new block structures. In addition, it considers the pole location of the block structure as a function of block length and relates it to the original pole location.

II. New Block Structures

Consider an M^{th} order IIR digital filter characterized by a transfer function given by

$$H(z) = \frac{z\{y_n\}}{z\{x_n\}} = \left(\sum_{i=o}^{M} a_i z^{-i}\right) / \left(\sum_{i=o}^{M} b_i z^{-i}\right) \qquad (1)$$

where $\{y_n\}$ and $\{x_n\}$ are the output and the input sequences, respectively. The transfer function can alternately be expressed in terms of the impulse response as

$$H(z) = \sum_{n=o}^{\infty} h_n z^{-n} \qquad (2)$$

The IIR digital filter characterized by the transfer function

$H'(z) = 1/(\sum_{i=o}^{M} b_i z^{-i})$ will be called the associated all-pole filter and its impulse response coefficients will be denoted by h'_n.

In terms of convolution, Eq. (1) can be written in basically two different forms. The first form relates the transfer function coefficients to the impulse response coefficients:

$$a_n = h_n * b_n \tag{3a}$$

$$= b_n * h_n \tag{3b}$$

The second form relates the transfer function coefficients to the input and output samples:

$$y_n * b_n = x_n * a_n \tag{4}$$

Writing Eq. (4) in matrix form and applying block partitioning with block length $L \geq M$, Burrus [2], [3] arrived at the block equation:

$$\underline{Y}_{k+1} = -\underline{B}_o^{-1} \underline{B}_1 \underline{Y}_k + \underline{B}_o^{-1} \underline{A}_o \underline{X}_{k+1} + \underline{B}_o^{-1} \underline{A}_1 \underline{X}_k \tag{5}$$

where $\underline{Y}_k$ and $\underline{X}_k$ are the k^{th} output and input vectors of length L:

$$\underline{Y}_k = [\, y_{kL} \; y_{kL+1} \; y_{kL+2} \; \cdots \; y_{(k+1)L-1}]^t$$

$$\underline{X}_k = [\, x_{kL} \; x_{kL+1} \; y_{kL+2} \; \cdots \; x_{(k+1)L-1}]^t$$

and, $\underline{A}_0$, $\underline{A}_1$, $\underline{B}_0$, and $\underline{B}_1$ are LxL matrices ($L \geq M$) given as:

$$\underline{A}_0 = \begin{bmatrix} a_0 & 0 & 0 & \cdots & 0 & 0 \\ a_1 & a_0 & 0 & \cdots & 0 & 0 \\ a_2 & a_1 & a_0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ a_M & a_{M-1} & a_{M-2} & \cdots & 0 & 0 \\ 0 & a_M & a_{M-1} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_1 & a_0 \end{bmatrix} \quad \underline{B}_0 = \begin{bmatrix} b_0 & 0 & 0 & \cdots & 0 & 0 \\ b_1 & b_0 & 0 & \cdots & 0 & 0 \\ b_2 & b_1 & b_0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ b_M & b_{M-1} & b_{M-2} & \cdots & 0 & 0 \\ 0 & b_M & b_{M-1} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & b_1 & b_0 \end{bmatrix}$$

$$\underline{B}_1 = \begin{bmatrix} 0 & 0 & \cdots & 0 & b_M & b_{M-1} & \cdots & b_1 \\ 0 & 0 & \cdots & 0 & 0 & b_M & \cdots & b_2 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & b_M \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$\underline{A}_1 = \begin{bmatrix} 0 & 0 & \cdots & 0 & a_M & a_{M-1} & \cdots & a_1 \\ 0 & 0 & \cdots & 0 & 0 & a_M & \cdots & a_2 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & a_M \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

In a similar manner, expressing Eq. (3a) in matrix form and applying block partitioning we arrive at:

$$\underline{A}_0 = \underline{H}_0 \, \underline{B}_0 \tag{6}$$

$$\underline{A}_1 = \underline{H}_1 \, \underline{B}_0 + \underline{H}_0 \, \underline{B}_1 \tag{7}$$

where $\underline{H}_0$ and $\underline{H}_1$ are LxL convolution matrices of the impulse responses of the original IIR filter:

$$\underline{H}_0 = \begin{bmatrix} h_0 & 0 & \cdots & 0 \\ h_1 & h_0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ h_{L-1} & h_{L-2} & \cdots & h_0 \end{bmatrix}, \qquad \underline{H}_1 = \begin{bmatrix} h_L & h_{L-1} & \cdots & h_1 \\ h_{L+1} & h_L & \cdots & h_2 \\ \vdots & \vdots & & \vdots \\ h_{2L-1} & h_{2L-2} & \cdots & h_L \end{bmatrix}$$

Applying the same procedure we obtain from Eq. (3b):

$$\underline{A}_0 = \underline{B}_0 \, \underline{H}_0 \tag{8}$$

$$\underline{A}_1 = \underline{B}_0 \, \underline{H}_1 + \underline{B}_1 \, \underline{H}_0 \tag{9}$$

From Eqns. (8) and (9) it follows that

$$\underline{B}_0^{-1} \, \underline{A}_0 = \underline{H}_0 \tag{10}$$

$$\underline{B}_0^{-1} \, \underline{A}_1 = \underline{H}_1 + \underline{B}_0^{-1} \, \underline{B}_1 \, \underline{H}_0 \tag{11}$$

Also, it can be shown that $\underline{B}_0^{-1} = \underline{H}'_0$, where $\underline{H}'_0$ is the convolution matrix of the associated all-pole filter, i.e.

$$\underline{H}'_0 = \begin{bmatrix} h'_0 & 0 & \cdots & 0 \\ h'_1 & h'_0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ h'_{L-1} & h'_{L-2} & \cdots & h'_0 \end{bmatrix}$$

The proof follows from Eq. (8) as for the associated all-pole filter, $\underline{A}_0$ is an unit matrix and $\underline{A}_1$, is a null matrix. Eqs. (8) and (9) were derived by Burrus.

Now, substituting Eqs. (10) and (11) in Eq. (5) we obtain

$$\underline{Y}_{k+1} = -\underline{B}_0^{-1} \underline{B}_1 \underline{Y}_k + \underline{H}_0 \underline{X}_{k+1} + \underline{H}_1 \underline{X}_k + \underline{B}_0^{-1} \underline{B}_1 \underline{H}_0 \underline{X}_k \tag{12}$$

Let us introduce a new vector $\hat{\underline{Y}}_k$ defined as:

$$\underline{Y}_k = \hat{\underline{Y}}_k + \underline{H}_0 \underline{X}_k \tag{13}$$

which when substituted in Eq. (12) yields:

$$\hat{\underline{Y}}_{k+1} = - \underline{B}_0^{-1} \underline{B}_1 \hat{\underline{Y}}_k + (-\underline{B}_0^{-1} \underline{B}_1 \underline{H}_0 + \underline{B}_0^{-1} \underline{A}_1) \underline{X}_k$$

Utilizing Eq. (11) in above we obtain

$$\hat{\underline{Y}}_{k+1} = - \underline{B}_0^{-1} \underline{B}_1 \hat{\underline{Y}}_k + \underline{H}_1 \underline{X}_k \tag{14}$$

Equations (14) and (13) form a new block implementation of the original IIR digital filter. This pair of equations are identical in form to the state-variable description of a multi-input, multi-output discrete-time system. Analytical solution for this set of equations can be written as:

$$\underline{Y}_k = (-\underline{B}_0^{-1} \underline{B}_1)^k \hat{\underline{Y}}_0 + \sum_{j=0}^{k-1} (-\underline{B}_0^{-1} \underline{B}_1)^{k-j-1} \underline{H}_1 \underline{X}_j + \underline{H}_0 \underline{X}_k \tag{15}$$

Another block implementation is obtained by realizing the IIR filter in two stages, i.e. realizing the poles and zeros separately. Thus, we can rewrite Eqn. (1) as a set of two equations:

$$\frac{z\{u_n\}}{z\{x_n\}} = 1 / (\sum_{i=0}^{M} b_i z^{-i}) \tag{16}$$

and

$$\frac{z\{y_n\}}{z\{u_n\}} = (\sum_{i=0}^{M} a_i z^{-i}) \tag{17}$$

where $\{u_n\}$ is the output of the pole-section. By expressing Eqns. (16) and (17) in matrix form and partitioning for a block length $L \geq M$ we arrive at

$$\underline{B}_0 \underline{U}_{k+1} + \underline{B}_1 \underline{U}_k = \underline{X}_{k+1} \tag{18}$$

$$\underline{Y}_k = \underline{A}_0 \underline{U}_k + \underline{A}_1 \underline{U}_{k-1} \tag{19}$$

where

$$\underline{U}_k = [u_{kL} \; u_{kL+1} \cdot \cdot \cdot \cdot \; u_{(k+1)L-1}]^t \tag{20}$$

Combining Eqns. (18) and (19) we obtain

$$\underline{Y}_k = (-\underline{A}_0 \underline{B}_0^{-1} \underline{B}_1 + \underline{A}_1)\underline{U}_{k-1} + \underline{A}_0 \underline{B}_0^{-1} \underline{X}_k \tag{21}$$

Making use of Eqns. (6) and (7) we rewrite above as:

$$\underline{Y}_k = \underline{H}_1 \underline{B}_0 \underline{U}_{k-1} + \underline{H}_0 \underline{X}_k \tag{22}$$

Eqn. (18) can be alternately written as:

$$\underline{U}_k = - \underline{B}_0^{-1} \underline{B}_1 \underline{U}_{k-1} + \underline{B}_0^{-1} \underline{X}_k \tag{23}$$

Eqns. (22) and (23) constitute a new block structure.

An alternate form of the above block description can be derived by defining a new variable

$$\underline{S}_k = \underline{B}_0 \underline{U}_k \tag{24}$$

in (23) which yields:

$$\underline{S}_k = - \underline{B}_1 \underline{B}_0^{-1} \underline{S}_{k-1} + \underline{X}_k \tag{25}$$

Similar substitution in (22) yields:

$$\underline{Y}_k = \underline{H}_1 \underline{S}_{k-1} + \underline{H}_0 \underline{X}_k \tag{26}$$

Eqns. (25) and (26) constitute the desired new block structure.

An entirely different block structure is obtained by putting the matrix representation of a digital filter advanced by Oppenheim and Crochiere [9] in block form. From the signal-flow-graph of the digital filter, the corresponding matrix representation is of the form:

$$\underline{W}_k = (\underline{I} - \underline{F}_c)^{-1} \underline{F}_d \underline{W}_{k-1} + (\underline{I} - \underline{F}_c)^{-1} \underline{b}\, x_k \tag{27}$$

$$y_k = \underline{C}\, \underline{W}_k \tag{28}$$

where $\underline{W}_k$ is a (N x 1) vector whose ith elements is the kth value of the signal sequence at mode i, $\underline{F}_c$ and $\underline{F}_d$ are N x N matrices, $\underline{b}$ is a N x 1 column vector, x_k is the kth member of the input sequence and y_k is the kth member of the output sequence. It is known that for a digital filter structure with no delay-free-loops, $\underline{F}_c$ is a lower triangular matrix with all zero main diagonal elements. This insures the existence of the inverse of $(\underline{I} - \underline{F}_c)$. Eqns. (27) and (28) can be easily put in block form as [7],

$$\begin{aligned} \underline{W}_{(k+1)L} &= \underline{A}^L \underline{W}_{kL} + \underline{B}\, \underline{X}_k \\ \underline{Y}_k &= \bar{C} \underline{W}_{kL} \end{aligned} \tag{29}$$

where,

$$\underline{A} = (\underline{I} - \underline{F}_c)^{-1} \underline{F}_d$$

$$\underline{B} = [\underline{A}^{L-1}\bar{b}, \underline{A}^{L-2}\bar{b}, \ldots \ldots \ldots \underline{A}\bar{b}\, , \bar{b}]$$

$$\bar{b} = (\underline{I} - \underline{F}_c)^{-1} \underline{b}$$

$$\bar{C} = [\underline{C}\, , \underline{C}\, \underline{A}, \underline{C}\, \underline{A}^2, \ldots \ldots \ldots \underline{C}\, \underline{A}^{L-1}]^t$$

$\underline{X}_k$- input block of length L

$\underline{Y}_k$- output block of length L

All these structures proposed here are identical in form to the state variable description of a multi-input, multi-output discrete-time system. For all of these sets analytical solution can be written like the one in equation (15).

Additional new block structures can be derived from each set of these newly proposed structures using linear transformation of the "State-variables". For example, if we define a new set of variables $\underline{V}_k$ via the transformation

$$\underline{V}_k = \underline{T}\,\underline{S}_k \qquad (30)$$

Where $\underline{T}$ is a non-singular L x L matrix, then from Eqns. (25) and (26) we arrive at

$$\underline{V}_k = -\,(\underline{T}\,\underline{B}_1\,\underline{B}_0^{-1}\,\underline{T}^{-1})\,\underline{V}_{k-1} + \underline{X}_k \qquad (31a)$$

$$\underline{Y}_k = \underline{H}_1\,\underline{T}^{-1}\,\underline{V}_{k-1} + \underline{H}_0\,\underline{X}_k \qquad (31b)$$

As can be seen from above, there are literally an infinite number of such block descriptions. However, it may be possible to choose a suitable transformation matrix $\underline{T}$ for a specific IIR filter such that an implementation of (31) would require the least amount of computation. For example, if we select $\underline{T}$ such that $\underline{B}_1\,\underline{B}_0^{-1}$ is transformed into Jordan canonical form, then the number of multiplications needed to compute $\underline{V}_k$ in (31a) is drastically reduced. At the same time, computation complexity in (31b) remains the same.

Examination of different implementations is important to exploit the computational advantage and also it is important because different structures exhibit different error characteristics due to finite length of the registers involved during implementation.

III. Pole Location of Block Structure

We first observe that in all of the block structures, the poles of block transfer matrices are given by the eigenvalues of the matrix $-\underline{B}_0^{-1}\,\underline{B}_1$. Burrus [2] conjunctored that these eigenvalues are given by the poles of the original sealar transfer function raised to the power L where L is the block length. We prove here that his conjuncture is indeed true. The proof is given in two parts i.e. first for L=M and then for any L ⩾M. It is simple to note that the eigenvalues of $-\underline{B}_0^{-1}\,\underline{B}_1$ are the same as that of $-\underline{B}_1\,\underline{B}_0^{-1}$.

Consider the case of L=M. Let us denote for this case $\underline{B}_1$ as $\underline{B}_1$ and $\underline{B}_0$ as $\underline{B}_0$. We shall show that in this case $-\underline{B}_1\,\underline{B}_0^{-1} = -\,\underline{B}_1\,\underline{B}_0^{-1} = \underline{A}^M$ where $\underline{A}$ is an M x M matrix given by,

$$\underline{A} = \begin{bmatrix} -b_1 & 1 & 0 & \cdots & 0 \\ -b_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -b_{M-1} & 0 & 0 & \cdots & 1 \\ -b_M & 0 & 0 & \cdots & 0 \end{bmatrix}$$

To prove our assertion it is sufficient to prove that

$$\underline{A}^M \, \underline{\mathcal{B}}_0 = - \underline{\mathcal{B}}_1 \tag{32}$$

In order to prove (32) we need the recurrence relation to compute $\underline{A}^P$ iteratively. Let us denote the $(i,j)^{th}$ element of $\underline{A}^P$ as a^P_{ij} now $\underline{A}^P$ can be expressed as

$$\underline{A}^P = \underline{A}^{P-1} \, \underline{A} \tag{33}$$

From (33) it can be shown that the elements of $\underline{A}^P$ are given interms of that of $\underline{A}^{P-1}$ as follows:

$$a^P_{i1} = - \sum_{j=1}^{M} b_j \, a^{P-1}_{ij} \tag{34}$$

$$a^P_{i,k} = a^{P-1}_{i,k-1} \tag{35}$$

with $a^1_{i,1} = - b_i$, $a^1_{ik} = 1$ for $k = i+1$ and $a^1_{ik} = 0$ otherwise for $i = 1,2,\ldots,M$, $k = 2,3,\ldots, M$. From (34) and (35) it can be seen that only the first column of $\underline{A}^P$ needs to be computed. All the others are obtained from $\underline{A}^{P-1}$ by shifting one column to the right. In otherwords, i^{th} column of $\underline{A}^P$ is $(i-1)^{st}$ column of $\underline{A}^{P-1}$. If we denote $\underline{A}^P_{ci}$ as i^{th} column of the matrix $\underline{A}^P$, then,

$$\begin{aligned} \underline{A}^P &= [\underline{A}^P_{C1}, \underline{A}^{P-1}_{C1}, \ldots, \underline{A}^1_{C1}, \underline{A}^1_{C2}, \ldots, \underline{A}^1_{CM-P+1}] \text{ for } P < M \\ &= [\underline{A}^P_{C1}, \underline{A}^{P-1}_{C1}, \ldots, \underline{A}^1_{C1}] \text{ for } P = M \\ &= \underline{A}^P_{C1}, \underline{A}^{P-1}_{C1}, \ldots, \underline{A}^{P-M+1}_{C1}] \text{ for } P > M \end{aligned} \tag{36}$$

From (33) and (36) it can be shown that,

$$\underline{A}^P_{C1} = - \underline{A}^{P-1} \, \underline{b}, \quad \text{where } \underline{b} = (b_1, b_2, \ldots b_M)^t. \tag{37}$$

It can be easily seen that $\underline{\mathcal{B}}_1$ is an upper triangular matrix, in which all the diagonal elements are identical. Let,

$$\underline{A}^M \, \underline{\mathcal{B}}_0 = \underline{C} \tag{38}$$

where $\underline{C}$ is an $M \times M$ matrix whose $(i,j)^{th}$ element is c_{ij}. To prove (32) all we need to show are the following:

(i) The elements of principal and upper diagonals of $\underline{C}$ are given by,

$$c_{j+1,\, j+k+1} = - b_{M-k}, \; j = 0, 1, \ldots, M-k-1, \; k = 0, 1, \ldots, M-1$$

(ii) All the lower triangular elements of $\underline{C}$ are identically equal to zero.

$$c_{j+k+2,\, j+1} = 0, \; j = 0, 1, \ldots, M-k-2, \; k = 0, 1, \ldots, M-2$$

Index j selects the elements in a given diagonal and k selects the specific diagonal.

Proof for (i): In this part unless otherwise specified, $j = 0, 1, \ldots M-k-1$ and $k = 0, 1, \ldots, M-1$. From (38),

$$c_{j+1,\, j+k+1} = \sum_{i=1}^{M} b_{i-j-k-1} \, a^M_{j+1,\, i}$$

with $b-i = 0$, $i = 1,2,\ldots,$ and $b_0 = 1$. Separating $i = (j+k+1)^{st}$ term from the above summation,

$$c_{j+1,\, j+k+1} = a^{M}_{j+1,\, j+k+1} + \sum_{i=1}^{M} b_{i-j-k-1}\, a^{M}_{j+1,\, i}, \qquad i \neq j+k+1 \tag{39}$$

from (35) and (34),

$$a^{M}_{j+1,\, j+k+1} = a^{M-j-k}_{j+1,\, 1} = - \sum_{i=1}^{M} b_i\, a^{M-k-j-1}_{j+1,\, i}$$

Substituting the above in (39) we get,

$$c_{j+1,\, j+k+1} = - \sum_{i=1}^{M} b_i\, a^{M-k-j-1}_{j+1,i} + \sum_{i=1}^{M} b_{i-j-k-1}\, a^{M}_{j+1,i},\ i \neq j+k+1 \tag{40}$$

In the second summation of (40), $b_{i-j-k-1} = 0$ for $i < j+k+1$ and also noting $i \neq j+k+1$ we get,

$$c_{j+1,\, j+k+1} = - \sum_{i=1}^{M} b_i\, a^{M-k-j-1}_{j+1,\, i} + \sum_{i=j+k+2}^{M} b_{i-j-k-1}\, a^{M}_{j+1,i} \tag{41}$$

Separating the term i=M-k from the first summation and using the change of variables i' = i-j-k-1 in the second summation, Eqn. (41) becomes,

$$\begin{aligned} c_{j+1,\, j+k+1} = {} & - b_{M-k}\, a^{M-k-j-1}_{j+1,\, M-k} - \sum_{i=1}^{M} b_i\, a^{M-k-j-1}_{j+1,i}\ ,\ i \neq M-k \\ & + \sum_{i=1}^{M-k-j-1} b_i\, a^{M-k-j-1}_{j+1,i} \end{aligned} \tag{42}$$

In the matrix $\underline{A}$, $a^{1}_{j+1,j+2} = 1$. In $\underline{A}^{M-k-j-1}$, $(M-k)^{th}$ column is nothing but $(j+2)^{nd}$ column of $\underline{A}$. Therefore, it follows that,

$$a^{1}_{j+1,\, j+2} = a^{M-k-j-1}_{j+1,\, M-k} = 1 \tag{43}$$

$$a^{M-k-j-1}_{j+1,i} = 0 \text{ for } i > M-k \text{ and } M-k-j \leqslant i < M-k$$

Substituting the above results in (42) we get the required result,

$$c_{j+1,j+k+1} = - b_{M-k}$$

Proof for (ii): In this part unless otherwise specified, j = 0,1,..., M-k-2 and k = 0,1,..., M-2. From (38),

$$c_{j+k+2,\, j+1} = \sum_{i=1}^{M} b_{i-j-1}\, a^{M}_{j+k+2,\, i}$$

Separating $i = (j+1)^{st}$ term from the above equation we arrive at

$$c_{j+k+2,\, j+1} = a^{M}_{j+k+2,\, j+1} + \sum_{i=1}^{M} b_{i-j-1}\, a^{M}_{j+k+2,i},\ i \neq j+1 \tag{44}$$

from (35) and (34),

$$a^{M}_{j+k+2,\, j+1} = a^{M-j}_{j+k+2,\, 1} = - \sum_{i=1}^{M} b_i\, a^{M-j-1}_{j+k+2,i}$$

Noting that $i \neq j+1$ and $b_{i-j-1} = 0$ for $i < j+1$ and substituting the above into (44) we get,

$$c_{j+k+2,\, j+1} = - \sum_{i=1}^{M} b_i\, a^{M-j-1}_{j+k+2,\, i} + \sum_{i=j+2}^{M} b_{i-j-1}\, a^{M}_{j+k+2,i} \tag{45}$$

using change of variable i' = i-j-1 and making use of (35) in the second summation of (45) we arrive at,

$$c_{j+k+2,\, j+1} = - \sum_{i=M-j}^{M} b_i\, a^{M-j-1}_{j+k+2,\, i}$$

For the same reasons used for (43),

$$a^{M-j-1}_{j+k+2,\ i} = 1 \text{ for } i = M+k+1$$
$$= 0 \text{ for } M-j \leqslant i < M+k+1 \text{ and } i \geqslant M+k+1$$

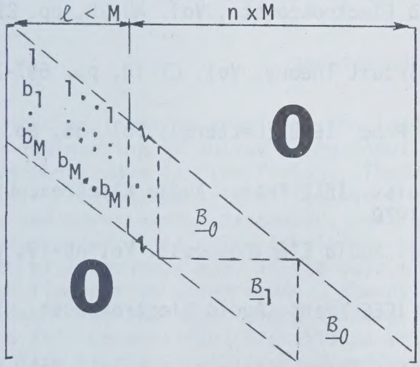
Substituting this in the above equation we arrive at the desired result,

$$c_{j+k+2,\ j+1} = 0.$$

Hence $\underline{A}^M \underline{B}_0 = - \underline{B}_1$. It is well known that eigenvalues of $\underline{A}$ are the poles of the original filter, since $\underline{A}$ is the companion matrix of the original difference equation. It is also a well-known result that eigenvalues of $\underline{A}^M$ are the eigenvalues of $\underline{A}$ raised to the power M. Hence the eigenvalues of $-\underline{B}_1 \underline{B}_0^{-1}$, which are the poles of the block difference equation, are the eigenvalues of the original scalar filter raised to the power L.

Finally for the case when $L \geqslant M$, the form of the matrices, $\underline{B}_0$ (the partition shown is purely for convenience) and $\underline{B}_1$ are as shown below:

$\underline{B}_0$ =



$$\underline{B}_1 = \begin{bmatrix} 0 & 0 & . & . & . & \underline{B}_1 \\ 0 & 0 & . & . & . & 0 \\ : & : & & & & : \\ 0 & 0 & . & . & . & 0 \end{bmatrix}$$

For simplicity let us assume that $L = 2M + \ell$, where $\ell < M$. Consider a L x L $\underline{D}$ matrix as given below, referring to (36),

$$\underline{D} = \begin{bmatrix} \underline{A}^{2M+\ell}_{c1}, & . & . & . & \underline{A}^{2M+1}_{c1}, & \underline{A}^{2M}_{c1}, & . & . & . & \underline{A}^{M+1}_{c1}, & \underline{A}^{M}_{c1} & . & . & . & \underline{A}^{1}_{c1} \\ & 0 & & & & 0 & & & & & & & 0 & \\ & 0 & & & & 0 & & & & & & & 0 & \end{bmatrix} \qquad (46)$$

Now it will be shown that $\underline{D}\,\underline{B}_0 = - \underline{B}_1$. This proof is based on the following identity,

$$\underline{A}^{(n+1)M} \underline{B}_0 + \underline{A}^{nM} \underline{B}_1 = [0] \qquad (47)$$

The proof of (47) directly follows from (32). Forming the product,

$$\underline{D}\,\underline{B}_0 = \begin{bmatrix} \underline{E}_{11} & \underline{E}_{12} & \underline{E}_{13} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

where $\underline{E}_{12}$ and $\underline{E}_{13}$ are M x M matrices and $\underline{E}_{11}$ is M x ℓ matrix, we get,

$$\underline{E}_{12} = \underline{A}^{2M} \underline{B}_0 + \underline{A}^{M} \underline{B}_1 = [0] \qquad \text{[from (47)]}$$

$$\underline{E}_{13} = \underline{A}^{M} \underline{B}_0 = - \underline{B}_1 \qquad \text{[from (32)]}$$

and making use of the equations (36) and (37) it can be shown that $\underline{E}_{11} = [0]$. Therefore $\underline{D}\,\underline{B}_0 = -\underline{B}_1$. Hence the eigenvalues of $-\underline{B}_0^{-1}\,\underline{B}_1$ are same as those of $\underline{D}$. The eigenvalues of $\underline{D}$ can easily be identified in terms of that of $\underline{A}$ by looking at the structure of $\underline{D}$. Referring to the structure of $\underline{D}$ in (46), it has $(M+\ell)$ zero eigenvalues and M eigenvalues of the matrix given below,

$$[\underline{A}_{c1}^{2M+\ell}, \underline{A}_{c1}^{2M+\ell-1}, \ldots \underline{A}_{c1}^{M+\ell+1}] = \underline{A}^{2M+\ell} \qquad \text{[from (36)]}$$

As we know the eigenvalues of $\underline{A}^{2M+\ell}$ $(=\underline{A}^{L})$ are that of original scalar filter raised to the power L. This proves Burrus Conjecture as stated before.

References

[1] C. S. Burrus, IEEE Trans. Audio Electroacoust., Vol. AU-20, pp. 230-235, Oct. 1972.

[2] ____________, IEEE Trans. on Circuit Theory, Vol. CT-18, pp. 697-701, Nov. 1971.

[3] B. Gold and K. L. Jordan, Jr., Proc. IEEE (Letters), Vol. 56, pp. 1717-1718, Oct. 1968.

[4] H. B. Voelcker and E. E. Hartquist, IEEE Trans. Audio Electroacoust. Vol. AU-18, pp. 169-176, June 1970.

[5] R. Read and J. Meek, IEEE Trans. Audio Electroacoust. Vol AU-19, pp. 322-323, Dec. 1971.

[6] J. W. Meek and A. S. Veletsos, IEEE Trans. Audio Electroacoust. Vol. AU-20, pp. 93-94, March 1972.

[7] R. A. Meyer and C. S. Burrus, IEEE Trans. Circuits & Syst. Vol. CAS-22, pp. 162-168, March 1975.

[8] __________________________, IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-24, pp. 53-58, Feb. 1976.

[9] A. V. Oppenheim and R. W. Schafer, Digital Signal Processing, Prentice Hall, Inc., New Jersey, 1974 (Chapter 4).

Acknowledgment

This work was supported by a National Science Foundation Grant ENG 74-08879.

STOCHASTIC POWER FLOW STUDY METHODS

G. T. HEYDT P. W. SAUER
Purdue Electric Power Center
Purdue University
West Lafayette, IN 47907

ABSTRACT

This paper endeavors to summarize the most recent efforts in accounting for uncertainties in power system load flow data. Different sources of uncertainty are characterized by different levels of uncertainty. The stochastic nature of the data gives rise to a statistical distribution for line power flow level and bus voltage profile. The paper goes on to focus on error in bus demand data due to inaccurate load forecast. The theory of multivariate normal random variables is discussed and applied to the stochastic power flow problem. The present status of stochastic power flow methods is discussed and conjectures are made about the future of the area.

BACKGROUND

Like most engineering problems, the load flow study problem in electric power engineering is solved using input data derived from certain physical measurements made in "the field". These measurements often contain uncertainty. Frequently, a load flow study is performed for some future time frame and power system parameters must be forecast--again the forecast data contains uncertainty. The levels of the uncertainty depend on the physical source of inaccuracy and, in the case of forecast parameters, the lead time to the time period under study. Recently, there have been several efforts to account for the uncertainty in load flow study input data rather than ignore this uncertainty. In this paper, several techniques for stochastic power flow studies are discussed and summarized. Some detail is presented in a method by which the probability of specific line loading and bus voltage levels is calculated.

The power flow problem is, in essence, the calculation of line loading and bus voltages of an electric power system from the generation/demand data and system topology and impedance information. Borkowska [1] is generally credited with recognizing the power flow problem as a stochastic problem due to potential inaccuracy of input data. Sources of this inaccuracy are discussed below. Dopazo et al. deserve recognition for recognizing the occurrence of "bad data" in power system parameters and for presenting a least squares detection and identification method to handle these "bad data" [2]. Following Borkowska in 1975 two additional methods for stochastic power flow calculations appeared in the literature [3,4]. Further evidence of interest in the area of stochastic studies in power engineering is the commissioning of the preparation of an extensive summary of probalistic methodologies for power systems [22]: this work was supported by the United States Energy Research Development Administration.

It is worthwhile to briefly diverge from this background summary to discuss the principal sources of uncertainty in power system planning and operation. Probalistic components occur in parameters of

i. The power system loads

ii. The power system interchange levels

iii. The system generation

iv. Possible outage of transmission system components

Each of these areas requires considerable attention in order to stochastically model the phenomena involved. For example, in system generation modelling, the probability distribution of output generation must be obtained by both experimental and theoretical studies. The validity of the distribution (i.e., to what machines does the model apply) must be clearly delineated. This area has been partially studied by Sufana and Heydt [5-9]. Furthermore the time period for which the stochastic model is valid for forecast errors in load levels, for example, is very different from that for measurement errors. The stochastic model of loads also has a nonstationary character since the demand differs in accuracy from season to season and even from day-of-the-week to day-of-the-week. In power system operation, a time frame in the near future is of interest and measurement errors and forecast errors due to weather are of salient importance. System planning, on the other hand, is done for a time frame much farther off into the future, and long term load forecasting error is the salient system demand error. Thus, the nature of the error is clearly dependent on the potential application: planning or operation.

POWER FLOW STUDIES

The basic power flow problem of determining bus voltages and line power flows based on a given load and generation schedule for a fixed network is a deterministic process. That is, one solution is obtained, without consideration for error in input data. Several fundamental problems in this area are: (a) accurately determining the statistical distributions of the random phenomena, (b) translation of general distributions of input variables (random bus loads) into usable distributions of output variables (random line power flows and bus voltages), (c) rapid and accurate evaluation of event probabilities.

A stochastic process is essentially a process involving random variables whose statistics are time dependent. The degree of time dependence is specified by the level of statistical stationarity. In this paper, the power system random inputs will be assumed to be stationary within specified time intervals. As such, the random variables exhibit statistical distributions which vary with time, but are assumed stationary over the time interval under study. The non-deterministic input data of this analysis are limited at this time to bus loads. That is, the system configuration (lines, transformers, etc.) are assumed fixed, and not subject to random failure. Further work in this area will include as many random features as is feasible. The random outputs of this analysis are limited at this time to line power flows. Since line power flows cannot be random quantities without simultaneous random bus voltages, there will be more than one non-deterministic output. The analysis of this paper focuses on the statistical distribution of line power flows resulting from random error in bus loads. In particular, the random input is attributed to error in load forecasting.

THE MULTIVARIATE GAUSSIAN DISTRIBUTION

At this point it is worthwhile to introduce the multivariate normal probability distribution. Because random errors can frequently be approximated by this model, some of the properties of this distribution are briefly described here. A vector on n dependent normal real random variables has a joint distribution which is the correlated multivariate normal distribution if its characteristic function is,

$$\Theta(T) = \exp(-\frac{1}{2}\ T^tCT)\exp(jT^tm)$$

where T, C, and m are the independent vector variable, convariance matrix, and mean vector respectively, and $j = \sqrt{-1}$. If the convariance matrix is non-singular, the corresponding density function is $f_X(X)$.

$$f_X(X) = (2\pi)^{-n/2}|C|^{-1/2}\exp(-\frac{1}{2}(X-m)^tC^{-1}(X-m)).$$

The multivariate normal distribution possesses a regenerative feature whereby any linear combination of the jointly normal random variables will also exhibit the multivariate normal distribution. That is, if elements of vector X, $X = (x_1 \cdots, x_n)^t$, are jointly normal where X has mean vector M and convariance matrix C, then Y = AX+B (where A is an mxn transformation matrix, and B is an m-vector of constants) will have multivariate normal distribution with mean vector AM+B and convariance matrix ACA^t. This property will form the basis for determining the statistical distribution of line power flows represented as a linear combination of bus powers.

It is important to clarify at this time, that the preceeding statements and equations apply only to real random variables. In general, complex random variables are treated as two random variables, although in certain cases, they may be treated as one [10].

The basis for the stochastic power flow at this stage is the utilization of the multivariate normal distribution. Justification for doing so stems from several rational, none of which individually is totally sufficient.

(a) The multivariate normal distribution is virtually the only statistical distribution which can be readily manipulated through nontrival transformations.

(b) A great deal of work has been done in multivariate normal area, thus providing accurate integral evaluations and "goodness of fit" methods for raw data.

(c) The central limit theorem prescribes that a large sum of certain independent random variables approach the normal distribution. Although this paper does not assume independence between bus loadings, the basic shape of the normal density function has often been observed in actual field data of random error.

(d) Errors, such as physical measurement errors, are frequently observed to be normal.

The reader should be warned, however, that the assumption of normal or near-normal statistics is a present bottleneck which must be overcome.

Dopezo et al. [3] avoided the problem of type of distribution of system demands by working only with the mean vector and the covarience matrix of those demands. The covarience and mean, however, do not specify a distribution and therefore can not be used for probability calculations. Future work in this area will be centered on verification of the normality assumption through actual data analysis, and goodness of fit testing.

Bus Voltage and Power Flow Linearization

In this study, random variations about a fixed operating point will be studied. The random variations will be taken as the real injected power at each system bus. In order to exploit the regenerative feature of the multivariate normal distribution, the relationship between line powers and bus powers must be linear. The linearized technique has been well documented by Heydt and others [4,11,12], and is written as

$$S_{line}^{B} = S_{line}^{A} + \rho \Delta S_{bus}$$

where S_{line}^{B} is the vector of line power flows under schedule, B, ΔS_{bus} is the vector of variations in bus powers about schedule A. ρ is an ℓxn matrix of complex distribution factors relating the ℓ line flows to the n bus injections. The ρ matrix is calculated as,

$$\rho_{line\ a,\ bus\ k} = \frac{(Z_{bus}^{S}(i,k) - Z_{bus}^{S}(j,k))^{*}}{z_a}$$

where line a joins busses i and j, z_a is the primitive line impedance. The elements of Z_{bus}^{S} are taken referenced to the swing bus of the system.

The second "output" of a power flow study other than line power loading is bus voltage magnitude, $|V_{bus}|$. $|V_{bus}|$ is related to S_{bus} through the Jacobian matrix [13,14]. Dopazo et al. [3] have shown how to write the $|V|$ expression analogous to the S_{line} expression above. The ρ matrix is simply replaced by the system Jacobian and S_{line} is replaced by $|V_{bus}|$.

Calculation of Probabilities

For a known correlated multivariate normal distribution, certain probabilities may be calculated easily. In particular, probabilities involving one randome variable requires only the error function evaluation,

$$P\{s_1 \geq u_1\} = \frac{1}{2} - \mathrm{erf}\left(\frac{u_1 - n_1}{\sigma_1}\right) ,$$

where n_1 and σ_1 are the mean and standard deviation of the normal variate respectively. For joint probabilities, a method illustrated by Kendall, known as the generalized tetrachoric series [15] provides a means to evaluate the multiple integral associated with joint probability evaluations for n variates,

$$P\{\underline{S} \geq \underline{U}\} = \int_{u_n}^{\infty} \cdots \int_{u_1}^{\infty} f_S(S) ds_1 \ldots ds_n$$

where f_S (S) is the joint density function of the multivariate normal random variables s_1 to s_n. The covariance matrix associated with variates s_1 to s_n is normalized to have a unity diagonal and correlation coefficients on the off diagonal. Call this matrix Q,

$$Q = \begin{bmatrix} 1 & q_{12} & q_{13} & \cdots q_{1\ell} \\ q_{12} & 1 & q_{23} & q_{2\ell} \\ q_{13} & q_{23} & 1 & q_{3\ell} \\ \cdot\ \cdot\ \cdot & & & \\ q_{1\ell} & q_{2\ell} & q_{3\ell} & 1 \end{bmatrix}$$

Construct a table of $\frac{1}{2}\ell(\ell - 1)$ entries which contains all off diagonal (non-unity) entries in the upper right triangle of Q; in this paper such a table is termed a Hermite Table [4] and consists of upper right triangle entries of Q versus index notations $n_1, n_2, \ldots n_{1/2\ell(\ell - 1)}$. For the Q shown above, the Hermite Table is,

Entry	Index
q_{12}	n_1
q_{13}	n_2
q_{14}	n_3
...	...
$q_{1\ell}$	$n_{\ell-1}$
q_{23}	n_{ℓ}
q_{24}	$n_{\ell+1}$
...	...
$q_{\ell-1,\ell}$	$n_{1/2\ell(\ell-1)}$

The required integral is evaluated as follows,

$$P(S > U = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \ldots \sum_{n_m=0}^{\infty} (n_1! n_2! \ldots n_m!)^{-1} .$$

$$. q_{12}^{n_1} q_{13}^{n_2} \ldots q_{\ell-1,\ell}^{n_m}) \ . \ (H_{n_1+n_2+n_3+\ldots+n_{\ell-1}-1}(u_1) .$$

$$. H_{n_1+n_\ell+n_{\ell+1}+\ldots-1}(u_2) H_{n_2+n_\ell+\ldots-1}(u_3) \ldots).$$

$$. (G(u_1) G(u_2) \ \ldots \ G(u_\ell)),$$

where

$$m = \frac{\ell(\ell-1)}{2}$$

and $H_r(x)$ is the r^{th} order Hermite polynomial of argument x. The order of each Hermite polynomial is obtained from the Hermite Table as follows: the first Hermite polynomial has a total order equal to minus one plus the sum of the indices corresponding to all entires which contain a "1" in the subscript of q. Thus in the Hermite Table, entries $q_{12}, q_{13}, q_{14}, \ldots q_{1\ell}$ each contain a "1" in the subscript of q and the corresponding indices are n_1, $n_2, \ldots n_{\ell-1}$. Therefore, the order of the first Hermite polynomial is

$$n_1 + n_2 + \ldots + n_{\ell-1} - 1.$$

The second Hermite polynomial has order found in the same way: entries q_{12}, $q_{23}, q_{24}, \ldots, q_{2\ell}$ each contain a "2" in the subscript of q. Corresponding indices in the Hermite table are n_1, $n_\ell, n_{\ell+1}, \ldots n_{2\ell-3}$, and the required order of the Hermite polynomial is

$$n_1 + n_\ell + n_{\ell+1} + \ldots + n_{2\ell-3} - 1.$$

All ℓ Hermite polynomials have order found in this way. The u_i terms are elements of vector U.

Heydt and Sauer [4,23] show several examples of the use of the method shown above. Note that although the explanation of the evaluation of the required $P(S > U)$ may appear formidable, digital implementation is not too difficult since the calculation of Hermite polynomial order involves fairly simple logic. Also, for small values of q, the series converges rapidly, and when correlations are less than 0.5, it is found that truncation of the infinite sums at about n = 20 will suffice. Also it has been found in actual implementations that usually the correlation coefficients in Q are either very near unity or zero. In either case, the generalized tetrachoric series method shown above should not be used: when the values of S_{line} (or $|V_{bus}|$) are uncorrelated (i.e., off diagonal entries of Q are near zero), the required probability is the product of several error functions since the random variables are independent [16-19]. When correlation coefficients are near unity, the assumption of perfect (unity) correlation is valid, and the required probability calculation is simply a single evaluation of an error function--the one corresponding to the least likely event in the joint probability [16-19].

THE PRESENT STATUS OF STOCHATIC LOAD FLOW STUDIES

The methods to calculate the probability of S_{line} and $|V_{bus}|$ within specified ranges are known when normal or near normal demand/generation data are assumed. It remains now to prove the method in practical implementation on historical data, thereby proving and demonstrating the value of the technique. Several governmental agencies are funding research in the area, and limited applications in the area of interchange power calculations using probabilistic techniques have been in existance for about one year [20-21]. The present status of stochastic power flow calculations is largely one of working closely with operating utilities to obtain system data and to obtain acceptance of the output data supplied by this technique.

The near future in this area surely will be occupied by continued applications. The streamlining of the digital implementation would be desirable since execution times still present a discouragement to potential users.

Considerable work remains in the area particularly of estimating load forecast error distributions, using goodness of fit methods to permit the use of the "best" Gaussian distribution to non-Gaussian data, and in incorporating bus voltage probability density functions to calculate probability of system failure. Also, the authors are now developing a second order correction technique to account for non-Gaussian, skewed densities.

SUMMARY AND CONCLUSIONS

The electric power flow problem may be viewed as a non-deterministic problem since load forecasting involves unavoidable uncertainty. Rather than ignore this uncertainty, it may be estimated and used in conjunction with linearized power flow methods to obtain bus voltage and line loading statistics. With certain approximations, the line loading probability density is calculable, thereby permitting the calculation of the probability of specific line loading conditions. The calculation of line loading conditions (eg., overloads, negative generation margin in a region, and other specific conditions), is accomplished using the generalized tetrachoric series or by simplified methods (when correlation coefficients are near unity or near zero).

The present status of stochastic power flow applications is one of continued testing and verification. Potential applications in operation require models for load forecast errors due to weather and measurement errors. Applications in system planning require models for long term load forecast error. In each area, the availability of probabilities of overload, negative generation margin, low bus voltage, and a variety of other quantities will make stochastic power flow study methods valuable to both the system operator and planner.

REFERENCES

[1] B. Borkowska, "Probabilisitic Load Flow", IEEE Transactions on Power Apparatus and Systems, PAS-93, No. 3, 1974, pp. 752-759.

[2] J. Dopazo, O. Klitin, A. Sasson, "State Estimation for Power Systems: Detection and Identification of Gross Measurement Errors", IEEE PICA 8th Annual Conference, Minneapolis, Minn., June, 1973, Publication 73 CHO 740-1 PWR, pp. 313-318.

[3] J. Dopazo, O. Klitin, and A. Sasson, "Stochastic Load Flows", IEEE Transactions on Power Apparatus and Systems, PAS-94, No. 2, 1975, pp. 299-309.

[4] G. T. Heydt, "Stochastic Power Flow Calculations", IEEE Paper No. A 75 530-6, Summer Meeting, San Francisco, CA. 1975.

[5] C. R. Sufana, G. T. Heydt and P. W. Sauer, "Stochastic Modeling of Power Systems Loads and Generation with Applications in Load Flow Studies", Purdue University, TR-EE-75-35, October 1975.

[6] G. Irisarri, "Power System Expansion Planning", Ph.D. Thesis, Purdue University, August 1975.

[7] G. T. Heydt, "An Intermediate Progress Summary for Research on Electric Power System Interchange Capacity Evaluation Using Stochastic Techniques", Report TR-EE-75-31, September 1975, Purdue University, West Lafayette, Indiana.

[8] G. T. Heydt, C. R. Sufana, "The Classification of Generation Availability Models by Types", 1976 Milwaukee Symposium on Automatic Computation and Control, April 1976.

[9] G. T. Heydt, C. R. Sufana, "Classes of Statistical Generation Models", Submitted to IEEE Power Apparatus and Systems, 1977 Winter Power Meeting.

[10] N. R. Goodman, "Statistical Analysis Based on a Certain Multivariate Complex Gaussian Distribution (An Introduction)", Annals of Mathematical Statistics, Vol. 34, 1963, p. 152.

[11] E. Hines and H. Limmer, "Rapid Load Flow Program Using Superposition", Proc. Power Systems Computation Conference 1, Queen Mary College, London, England, 1963.

[12] W. M. Grady and G. T. Heydt, "Fast Three Phase Load Flow Studies", Technical Report TR-EE-72-28, Purdue University, West Lafayette, Indiana, December 1972.

[13] G. Stagg, A. El-Abiad, "Computer Methods in Power System Analysis", McGraw Hill Book Company, 1968.

[14] H. E. Brown, "Solution of Large Network, by Matrix Methods", J. Wiley Book Company, 1975.

[15] M. G. Kendall, "Proof of Relations Connected with the Tetrachoric Series and its Generalization", Biometrika -A Journal for the Statistical Study of Biological Problems, University Press, Cambridge, England, 1941-1942.

[16] H. Cramer, "Mathematical Methods of Statistics", Princeton University Press, Princeton, New Jersey, 1954.

[17] A. Papoulis, "Probability, Random Variables and Stochastic Processes" McGraw Hill, 1965.

[18] K. S. Miller, "Multidimensional Gaussian Distributions", John Wiley and Sons, New York, N. Y., 1964.

[19] H. Cramer, "The Elements of Probability Theory and Some of its Applications", John Wiley and Sons, New York, N. Y., 1955.

[20] G. T. Heydt, and B. M. Katz, "A Stochastic Model in Simultaneous Interchange Capacity Calculations", Vol. PAS-94, No. 2, March/April, 1975, pp. 350-359.

[21] S. W. Anderson, R. K. Angel, G. T. Heydt, B. M. Katz, G. L. Landgren, H. L. Terhune, "Simultaneous Power Interchange Calculations", Technical Report, TR-EE-74-1, January 1974, Purdue University, West Lafayette, Indiana.

[22] N. Reppen, R. Ringlee, B. Wollenberg, A. Wood, "Probabilisitc Methodologies--A Critical Review", 1975 ERDA Conference on Power System, Henniker, N. H.

[23] P. W. Sauer, G. T. Heydt, "Stochastic Power Flow--Line Power Calculations", 1976 Midwest Power Symposium, October 1976, Manhattan, Kansas.

NONLINEAR RESISTIVE CIRCUIT MODELS FOR POWER SYSTEM STEADY-STATE ANALYSIS

FELIX F. WU
University of California
Berkeley, California 94720

R. L. SULLIVAN
Energy Research and Development
Administration
Washington, D.C. 20545

ABSTRACT

The use of nonlinear resistive circuit models for power system steady-state design is examined. Two coupled models are developed which in one case relates real power and voltage phase angles, and in the other case relates the imaginary part of currents and the real part of voltages. The use of the models is demonstrated by considering the practical problem of planning reactive power compensation. For simplicity, a decoupled form of the models is used.

INTRODUCTION

Nonlinear resistive circuit theory evolved from the need for better analysis tools in designing electronic circuits. [1-4] Only recently have some of the salient features been applied to analyzing large power network problems. The appeal of these concepts most certainly stems from their use in automated electronic circuit design and their potential attractiveness in automated power network design. Invariably, in automated design one is faced with maximizing some design objective which requires knowledge of the objective function sensitivity to changes in the design parameters. The initial application of nonlinear resistive circuit theory to power system analysis was in the area of automated network design. [5] Although this initial effort employed a linearized model of the power network, it did open up a research area which has yet to run its course. Other investigators have attempted to formulate the automated power network design problem incorporating the more realistic nonlinear steady-state model of the network. [6] In so doing, complex network elements are introduced which are not strictly resistive.

As used in this paper, nonlinear resistive circuits are defined by a graph and each branch by its associated voltage and current. The branch voltages and currents satisfy Kirchhoff's voltage and current laws and are related in a nonlinear fashion. The inputs are the independent voltage and current sources. In contrast, the steady-state branch relations of power networks are linear but the performance equations are nonlinear since they describe not the voltage and currents throughout the network but rather the voltages and real and reactive power. Hence, power network problems are circuit problems and they are nonlinear, but they are not strictly speaking nonlinear circuit problems.

POWER SYSTEM STEADY-STATE MODEL AND ANALYSIS [7,8]

An electric power system contains generators and loads, which are interconnected by transmission lines and transformers. Transmission lines and transformers are usually modeled as lumped, linear, time-invariant RLC elements. The interconnection thus becomes a linear time-invariant RLC network. During steady-state operation, a generator supplies a specified real power and maintains a specified voltage magnitude. From a circuit-theoretic point of view, a generator can be viewed as an independent real-power-and-voltage-magnitude source. A load represents a specified real and reactive power demand which can be viewed as a negative independent real-and-reactive-power source. Since the total power supplied must be equal to the total power consumed plus the losses and the losses are not known a priori, one generator is assigned to take up the losses in the system. It is also customary to specify the voltage magnitude of this generator and to take the phase angle of this generator as the reference. Thus, this generator can be viewed as an independent voltage-magnitude-and-phase-angle source. Without loss of generality, let us assume that each node has only one source. Hence we may partition the nodes into three types: (i) PV-node, for which the real power input and the voltage magnitude are specified; (ii) PQ-node, for which the real and the reactive power inputs are specified; (iii) the slack-node, for which the voltage magnitude and the phase angle are specified. Consider a power system having N_v PV-nodes and N_q PQ-nodes. Let $Y_\beta = G_\beta + jB_\beta$ be the primitive admittances of the N_b elements and let E_β and I_β denote the complex voltage across and current through each element. Then for each element

$$Y_\beta E_\beta = I_\beta \qquad \beta=1\ldots N_b \tag{1}$$

From nodal analysis and the element model in eq. (1), the k^{th} complex node current, J_k, can be expressed as a function of the node voltages

$$\sum_{m=o}^{N} Y_{km} E_m = J_k \tag{2}$$

where node m=o is the slack node. Multiplying eq. (2) by the voltage conjugate, E_k^*, and separating into a real part P_k and an imaginary part Q_k, we obtain the so-called load-flow equations

$$\begin{aligned} P_k &= R(E_k^* \Sigma Y_{km} E_m) \\ Q_k &= -I(E_k^* \Sigma Y_{km} E_m) \end{aligned} \tag{3}$$

where R denotes "the real part" and I "the imaginary part." For simplicity, let all specified quantities be identified by ∿. For PQ-nodes we get the following model:

$$\begin{aligned} \tilde{P}_k &= R(E_k^* \sum_m Y_{km} E_m) \\ \tilde{Q}_k &= -I(E_k^* \sum_m Y_{km} E_m) \qquad k = N_v+1\ldots N \end{aligned} \tag{4}$$

Similarly for PV-nodes we have

$$\begin{aligned} \tilde{P}_k &= R(V_k e^{-j\theta_k} \sum_m Y_{km} E_m) \qquad k=1\ldots N_v \\ \tilde{V}_k &= |\tilde{E}_k| \end{aligned} \tag{5}$$

Finally for the slack node

$$
\begin{aligned}
P_k &= \mathcal{R}(\tilde{E}_k^* \sum_m Y_{km} E_m) \qquad k=1 \\
Q_k &= -\mathcal{I}(\tilde{E}_k^* \sum_m Y_{km} E_m)
\end{aligned}
\tag{6}
$$

Solving the load-flow problem basically involves solving simultaneously the 2N equations given in eqs. 4-5 for the 2N unknown variables. The slack node powers are calculated as the last step in the solution process.

Although we have just described the most common formulation of the load-flow problem, it is obvious that a different set of variables could be specified as known, and hence, a different set of variables become the unknowns. A straightforward but unrealistic formulation would be to assume all the voltages are specified and the 2N power injections are to be calculated. In some cases, especially when attempting to decouple the equations other formulations can be employed. [9-11] For our purposes here we will elect to use the more common formulation as described above.

NONLINEAR RESISTIVE CIRCUIT MODELS

To obtain the desired nonlinear resistive circuits $\mathcal{N}_1$ and $\mathcal{N}_2$, we will make the following assumption about the power network, denoted by $\mathcal{N}$, to be modeled: All the elements of $\mathcal{N}$ are lossless, and hence, the network consists of only inductive and capacitive elements. Because of the typically low resistance to inductance ratios associated with modern extra high voltage systems, this assumption is not particularly restrictive.

To develop the nonlinear resistive circuit models to describe the power network, we will proceed along two major fronts. First, we will investigate how the power network circuit elements may be characterized as nonlinear resistive elements. Then we will investigate how the generating PV-nodes and consuming PQ-nodes can be characterized by nonlinear resistive elements. The emphasis will be placed on attempting to define two nonlinear resistive networks that may replace the standard power-systems model for use in conjunction with Tellegen's theorem in developing automated design procedures. Since Tellegen's theorem is to be used the network variables chosen to describe the new networks must satisfy Kirchhoff's laws.

Consider a typical element β which connects nodes k and m, where m can be the datum. Let I_β, P_β, E_β, and ψ_β denote the current and power through and the voltage an angle across element β. Then clearly, the sum of ψ_β or E_β around any loop is zero, hence, they satisfy Kirchhoff's voltage law. Similarly, with losses ignored, P_β and I_β satisfy Kirchhoff's current law. P_β and ψ_β are the real power variables in the new network models, and $\mathcal{R}(E_\beta)$ and $\mathcal{I}(I_\beta)$ are the reactive power variables. From eq. (1) it is straightforward to show that for a lossless element connecting nodes k and m that

$$
P_\beta = \frac{\mathcal{R}(E_k)\mathcal{R}(E_m)B_\beta}{\operatorname{Cos}\theta_k \operatorname{Cos}\theta_m} \operatorname{Sin}(\psi_\beta = \theta_k - \theta_m) \tag{7}
$$

Also from eq. 1

$$
\begin{aligned}
\mathcal{I}(I_\beta) &= B_\beta \mathcal{R}(E_\beta) \\
&= B_\beta(\mathcal{R}(E_k) - \mathcal{R}(E_m))
\end{aligned}
\tag{8}
$$

As desired, eqs. (7) and (8) constitute the nonlinear resistive expressions needed to model the network elements. To develop the models for the PQ-nodes which constitute elements from the node in question to the datum, we have from eqs. (2) and (3)

$$\mathcal{I}(J_k) = -\mathcal{I}\left(\frac{\tilde{P}_k + j\tilde{Q}_k}{E_k^*}\right) = \frac{-\tilde{Q}_k \mathrm{Cos}^2\theta_k}{\mathcal{R}(E_k)} + \frac{\tilde{P}_k \mathrm{Sin}\theta_k \mathrm{Cos}\theta_k}{\mathcal{R}(E_k)} \tag{9}$$

and

$$P_k = \frac{\mathcal{R}(E_k)}{\mathrm{Cos}\theta_k} \sum_m V_m B_{km} \mathrm{Sin}(\theta_k - \theta_m) \tag{10}$$

Finally, for PV-nodes

$$\mathcal{I}(J_k) = \frac{Q_k \mathrm{Cos}\theta_k}{\tilde{V}_k} + \frac{\tilde{P}_k \mathrm{Sin}\theta_k}{\tilde{V}_k} \tag{11}$$

$$\tilde{P}_k = \tilde{V}_k \sum_m V_m B_{km} \mathrm{Sin}(\theta_k - \theta_m) \tag{12}$$

In summary, eqs. 7,10, and 12 describe the so-called real power nonlinear resistive circuit and eqs. 8, 9, and 11 describe the twin reactive power nonlinear resistive circuit. Clearly, the equations and circuits are coupled. In the special case when the coupling is weak as it is when the angles are small and the voltage magnitudes near unity the network models reduce to those given in Table 1.

Table 1

	Real Power Model (N_1)	Reactive Power Model (N_2)
elements	$P_\beta \simeq B_\beta \psi_\beta$	$\mathcal{I}(I_\beta) \simeq B_\beta \mathcal{R}(E_\beta)$
PV-nodes	$\tilde{P}_k \simeq \sum_m B_{km}(\theta_k - \theta_m)$	$\mathcal{I}(J_k) \simeq \frac{Q_k}{\tilde{V}_k}$
PQ-nodes	$\tilde{P}_k \simeq \sum_m B_{km}(\theta_k - \theta_m)$	$\mathcal{I}(J_k) \simeq \tilde{Q}_k$

Note that the real power nonlinear resistive model is now completely independent of the reactive power model, and thus, the two circuits may be solved independently. In fact, the real power model defined above constitute what has become known as the DC power flow model which was used, as explained earlier, in the initial attempts to automate power network planning. In view of the earlier work involving this model, the next section will be devoted to a discussion of how the reactive power nonlinear resistive circuit model can be used in what is called reactive power compensation planning.

REACTIVE POWER COMPENSATION PLANNING

In planning the expansion of a modern electric utility, it is quite common to be confronted with reactive power and voltage profile problems. [12-13] Typically during heavy load periods system voltages are depressed requiring in some cases the addition of capacitive compensation. Similarly, during light load periods system voltages tend to become too high, and especially in those systems that contain excessive amounts of EHV (Extra High Voltage) and underground cables, thus necessitating the use of inductive compensation. To determine a viable expansion strategy for reactive power compensation, it is possible to formulate the compensation problem as a nonlinear optimization problems in which the objective is often to minimize

$$E = \tfrac{1}{2} \sum_{k\epsilon\Omega_1} W_k (R(E_k) - R(E_{ko}))^2 \qquad (13)$$

w.r.t. the susceptances B_β of the compensation elements. [14]. Ω_1 denotes the set of all node voltages to be considered in the optimization process. $R(E_{ko})$ is the desired value of $R(E_k)$. Differentiating eq.11 w.r.t. the susceptances, $B_\beta, \beta\epsilon\Omega_2$, of the to be added shunt compensators, we get

$$\frac{\partial E}{\partial B_\beta} = \sum_{k\epsilon\Omega_1} W_k (R(E_k) - R(E_{ko})) \frac{\partial R(E_k)}{\partial B_\beta} \qquad \beta\epsilon\Omega_2 \qquad (14)$$

where $R(E_k)$ and $\frac{\partial R(E_k)}{\partial B_\beta}$ must be evaluated at the point corresponding to the solution of the reactive power nonlinear resistive circuit. Using the results of eq. (14), the susceptances of those compensators which are most effective in providing the compensation can be sequentially updated until the objective function is minimized. Let Ω_3 denote the set of all RLC networks elements and Ω_4 all PV elements from a node to the datum. Based on Tellegen's Theorem [15]

$$\sum_{\beta\epsilon(\Omega_1 \ldots \Omega_4)} R(\hat{E}_\beta)\Delta I(I_\beta) - I(\hat{I}_\beta)\Delta R(E_\beta) = 0 \qquad (15)$$

The cap denotes the quantities pertaining to the as yet undefined adjoint network. All other quantities are associated with the reactive power nonlinear resistive model developed earlier.

Assuming decoupling for simplicity and using the model in Table 1, for $\beta\varepsilon\Omega_3$

$$\mathcal{I}(I_\beta) = B_\beta \mathcal{R}(E_\beta)$$

or $$\Delta\mathcal{I}(I_\beta) = B_\beta \Delta\mathcal{R}(E_\beta) \tag{16}$$

Substituting eq. 16 into 15, we get

$$\sum_{\beta\varepsilon\Omega_3} (\mathcal{R}(\hat{E}_\beta)B_\beta - \mathcal{I}(\hat{I}_\beta))\ \Delta\mathcal{R}(E_\beta) \tag{17}$$

For $\beta\varepsilon\Omega_4$

$$\mathcal{I}(J_\beta) = \frac{Q_\beta}{\tilde{V}_\beta}$$

or $$\Delta\mathcal{I}(J_\beta) = \frac{\Delta Q_\beta}{\tilde{V}_\beta} \tag{18}$$

Substituting eq. 18 into eq. 15, we get

$$\sum_{\beta\varepsilon\Omega_4} \frac{\mathcal{R}(\hat{E}_\beta)}{\tilde{V}_\beta}\ \Delta Q_\beta \tag{19}$$

For $\beta\varepsilon\Omega_1$

$$\mathcal{I}(I_\beta) = -\gamma_\beta$$

or $$\Delta\mathcal{I}(J_\beta) = 0 \tag{20}$$

Substituting eq. 20 into eq. 15

$$\sum_{\beta\varepsilon\Omega_1} - \mathcal{I}(\hat{I}_\beta)\Delta\mathcal{R}(E_\beta) \tag{21}$$

For $\beta\varepsilon\Omega_2$

$$\mathcal{I}(I_\beta) = B_\beta \mathcal{R}(E_\beta)$$

or $$\Delta\mathcal{I}(I_\beta) = \mathcal{R}(E_\beta)\ \Delta B_\beta + B_\beta \Delta\mathcal{R}(E_\beta) \tag{22}$$

Substituting eq. 22 into eq. 15

$$\sum_{\beta\varepsilon\Omega_2} \mathcal{R}(\hat{E}_\beta)\mathcal{R}(E_\beta)\Delta B_\beta + (\mathcal{R}(\hat{E}_\beta)B_\beta - \mathcal{I}(\hat{I}_\beta))\Delta\mathcal{R}(E_\beta) \tag{23}$$

By letting

$$\mathcal{R}(\hat{E}_\beta)B_\beta = \mathcal{I}(\hat{I}_\beta) \qquad \beta\varepsilon(\Omega_3 + \Omega_2) \tag{24a}$$

$$\mathcal{R}(\hat{E}_\beta) = 0 \qquad \beta\varepsilon\Omega_4 \tag{24b}$$

$$\mathcal{I}(\hat{I}_\beta) = W_\beta(\mathcal{R}(E_\beta) - \mathcal{R}(E_{\beta o})) \qquad \beta\varepsilon\Omega_1 \tag{24c}$$

eq. 15 reduces to

$$\sum_{\beta\varepsilon\Omega_1} W_\beta(\mathcal{R}(E_\beta) - \mathcal{R}(E_{\beta o}))\Delta\mathcal{R}(E_\beta) = \sum_{\beta\varepsilon\Omega_2} \mathcal{R}(E_\beta)\mathcal{R}(\hat{E}_\beta)\Delta B_\beta \tag{25}$$

For a change in a particular compensator susceptance $\Delta B_\beta = \Delta B_i$, $\beta = i$, eq. 25 becomes

$$\sum_{\beta\varepsilon\Omega_1} W_\beta(\mathcal{R}(E_\beta) - \mathcal{R}(E_{\beta o}))\,\frac{\Delta\mathcal{R}(E_\beta)}{\Delta B_i} = \mathcal{R}(E_i)\mathcal{R}(\hat{E}_i)$$

Thus, eq. 14 for $\beta = i$ can be written as

$$\frac{\partial E}{\partial B_i} = \mathcal{R}(E_i)\mathcal{R}(\hat{E}_i) \qquad \text{all} \quad i\varepsilon\Omega_2 \tag{26}$$

In summary, by solving the reactive power nonlinear resistive network whose element relations are described by eq. 24 with all PV-nodes short circuited, eq. 24b, and all other nodes injected with currents defined in eq. 24c, the adjoint voltages, $\mathcal{R}(\hat{E}_i)$, can be determined and used along with the original circuit voltages $\mathcal{R}(E_i)$, $i\varepsilon\Omega_1$, to calculate the gradient of the objective function given in eq. 26 w.r.t. the susceptances of prospective compensators. Knowing the gradient, the susceptances of the most effective compensators can be updated until the objective function approaches zero, at which time the independent current sources driving the adjoint network approach zero as seen in eq. 24c.

CONCLUSION

We have developed two nonlinear resistive circuit models for use in automated power system network design. The first model when subjected to decoupling assumptions reduces to the well-known DC power-flow model used in early automated design procedures. The latter model when subjected to the same decoupling assumptions reduced to a very useful DC reactive power model. The usefulness of the nonlinear resistive reactive power model was demonstrated by showing that reactive power compensation planning can easily be formulated using Tellegen's Theorem as a nonlinear optimization problem. The gradient of the performance objective was shown to be easily determined using only the real parts of the original circuit voltages in conjunction with the adjoint voltages calculated using the developed reactive power nonlinear resistive model.

REFERENCES

1. S. W. Director and R. A. Rohrer, "Automated Network Design-the Frequency-Domain Case," IEEE Transactions on Circuit Theory, Vol. CT-16, pp. 331-337, August 1969.

2. L. O. Chua, Introduction to Nonlinear Network Theory, McGraw Hill, New York, 1969.

3. C. A. Desoer and E. S. Kuh, Basic Circuit Theory, McGraw Hill, New York, 1969.

4. L. O. Chua and P. M. Lin, Computer-Aided Analysis of Electric Circuits, Prentice-Hall, Englewood Cliffs, N.J., 1975.

5. R. Fischl and W. R. Puntel, "Computer Aided Design of Electric Power Transmission Networks," IEEE paper C72 168-8, 1972.

6. H. B. Püttgen and R. L. Sullivan, A Novel Comprehensive Approach to Power System Sensitivity Analysis," to be presented at Winter Power Meeting, 1977.

7. O. I. Elgerd, Electric Energy Systems Theory, McGraw Hill, New York, 1971.

8. G. W. Stagg and A. H. El-Abiad, Computer Methods in Power System Analysis, McGraw Hill, New York, 1968.

9. B. Stott, "Decoupled Newton Load Flow," IEEE Transactions on Power Apparatus and Systems, Vol. PAS-91, pp. 1955-1959, September 1972.

10. B. Stott and O. Alsac, "Fast Decoupled Load Flow," IEEE Transactions on Power Apparatus and Systems, Vol. PAS-93, pp. 859-869, May 1974.

11. F. F. Wu, "Theoretical Study of the Convergence of the Fast Decoupled Load Flow," to be published.

12. F. F. Wu, "Multiple Single-Element Modification of Large Scale Linear Networks," to appear in Recent Contributions to Large Scale Systems, edited by R. Saeks.

13. H. B. Puttgen, R. L. Sullivan, and J. Morf, "Economically Optimal Allocation of Reactive Power Sources Using a C.A.D. Technique," PSCC Proceedings, London, England, 1975.

14. R. R. Shoults and M. S. Chen, "Reactive Power Control by Least Squares Minimization," IEEE paper F75 426-7, presented at Power Engineering Society Summer Meeting, 1975.

15. E. Polak, Computational Methods in Optimization: A Unified Approach, Academic Press, New York, 1971.

16. S. W. Director and R. A. Rohrer, "The Generalized Adjoint Network and Network Sensitivities," IEEE Transactions on Circuit Theory, Vol. C7-16, pp. 318-323, 1969.

THE POTENTIAL FOR APPLICATION OF LINEAR AND QUADRATIC PROGRAMS IN SOLVING ECONOMIC DISPATCH AND EXPANSION PLANNING PROBLEMS

GYANESHWAR DAYAL
Harza Engineering Company
Chicago, Illinois

ABSTRACT

This paper presents results which show great potential for using linear and quadratic programs for solving economic dispatch and system expansion planning problems. It briefly discusses the recently developed re-inversion routine and procedures, for updating the basis inverse, which contribute to considerable saving in computational effort and computer memory requirements.

I. INTRODUCTION

In the past two decades, considerable research effort has been directed toward solving the complex problems of optimal load flow and power system expansion planning. Because of the size of the modern integrated power systems, importance of these problems can hardly be overemphasized. Many mathematical optimization techniques such as nonlinear, quadratic, linear, and dynamic programming have been used for solving these problems.

A general load flow optimization problem is defined as the one of selecting an operating state - from other infinite possibilities - which optimizes some power system quantity, without violating any of the problem constraints. A number of problems can be defined by different choices of the objective function and constraints. When the objective function is fuel cost, it is called the Economic Dispatch problem. The classical approach of solving this problem is equal incremental loading with system losses taken into account by using the B-constants method. In early 1960's, this classical approach was implemented for on-line economic dispatch. Because of its simplicity, this method is still being used by most of the utilities. The method works with a reduced set of variables and so it is incapable of handling problems involving security constraints such as bus angle spread, line current limits, and bus voltage inequality constraints. Consideration of security constraints requires more rigorous models for system representation. This results in considerable increase in the problem size. Some of the real world large size problems may have matrices of the order of 4000 or more. Real-time solution of such problems requires use of some of the most sophisticated sparsity techniques and elegant mathematical optimization methods.

Many non-linear programming techniques have been used for solving the economic dispatch problem. The most successful among them have been the Generalized Reduced Gradient and Hessian Matrix methods. In recent years, considerable interest has developed[1-6] in using mathematical optimization procedures which utilize linear and quadratic programs. Central Electricity Generating Board of England[6] has been pioneer in the application of linear programming approach for determining the most economical generation schedule, consistent with the network security requirements. More recently, Wollenberg and Stadlin[5] have shown that Dantzig-Wolfe's decomposition technique of linear programming can be very effectively used for real-time security dispatch. This requires linear or piecewise linear objective function subject to linear equality and inequality constraints. Nicholson and Sterling[3], Nabona and Freris[4] used Beale's quadratic programming approach for solving economic dispatch problem. Then Reid and Hasdorff[2] successfully used, for the first time, Wolfe's quadratic programming method for optimum active and reactive power dispatch using rigorous model for power system representation. The above mentioned quadratic programming methods have the basic potential for handling large size real-world problems. This is primarily because they require only minor modifications to the basic simplex algorithm. At present, highly efficient and sophisticated linear programming packages are commercially available. These codes have been extensively used, in the industry, for solving problems having as many as a few thousand constraints.

Power system expansion planning is another area, of interest, in which mathematical optimization techniques are increasingly being used. The expansion planning work has mainly been done using the simulation approach. This involves postulation and cost estimation of a large number of feasible expansion strategies. The computational effort required to evaluate all the proposed alternatives is exceedingly high. Moreover, in spite of the simulation of so many plans, there is no assurance that the optimum plan will be among them. The abovesaid difficulties can be overcome by using the mathematical optimization techniques. These techniques automatically develop an optimum expansion pattern. The optimum expansion planning problem is defined as that of minimizing the present worth of future costs, both capital and operating, subject to meeting system operating and reliability requirements. The expansion planning problem is mathematically very complex and large. This is due to some of the inherent nonlinear constraints and generally the time-phased nature of the optimization problem. Various optimization techniques - such as dynamic programming, quadratic and linear programming - have been used for solving this problem. Linear programming approach has received much attention[7-10] in respect to this problem. However, it is noticed that the planning problem inherently has some of the variables which are constrained to have only integer values. This further complicates the problem. Mixed-integer programming approach has been found to be more effective for handling such problems. In this, a continuous optimal solution is first found using regular linear programming technique. Then "branch-and-bound" method is used, beginning with the continuous solution, to find the optimal solution respecting the integer requirements of the specified variables.

Reid and Hasdorff[2] used Rand Corporation's quadratic programming package RSQPF4 for solving economic dispatch problem. However, it was observed that this computer code could not handle, very efficiently, large size economic dispatch problems. The storage requirements, for the basis inverse, were found to increase rather rapidly with increase in the problem size. They reported that, in the case of IEEE 57 bus system, the maximum number of nonzero elements in the basis inverse were about 10,000. However, for the 118 bus system, these nonzero elements jumped to about 36,000. So it was concluded that the problem of excessive storage requirements will have to be tackled before Wolfe's quadratic programming method can be used for solving real large size economic dispatch problems.

The primary objective of this research has been to develop and solve various techniques that can be employed for overcoming the abovesaid difficulties. Recently, Forrest and Tomlin[11] have developed highly efficient procedures for updating the triangular factors at each iteration of the simplex method. It has been shown that the power of the simplex method can further be greatly increased if Forrest and Tomlin's algorithm is used in conjunction with the extremely efficient new inversion technique -advocated by Hellerman and Rarick[12].

The abovementioned techniques have been successfully implemented in the new computer code developed during the course of this research. This code was used for solving economic dispatch problems. Results of the experiment and salient points of some of the newly developed procedures are discussed in the following sections.

II. TECHNIQUES FOR CONTROLLING STORAGE REQUIREMENTS OF THE BASIS INVERSE

The commercially available linear programming packages have already reached a very high level of sophistication and efficiency. However, considerable research effort is still being directed toward developing more efficient algorithms which can help in keeping storage requirements within reasonable bounds. Some of these effective techniques are presented below.

2.1 Forrest and Tomlin's Method of Updating Basis Inverse

Forrest and Tomlin have developed a new method[11] for updating the triangular factors at each iteration of the simplex method. This new algorithm has proved to be very effective in reducing the overall length of the file carrying the basis inverse. Tests on medium to large size representative practical problems have shown that with this approach the growth rate of new nonzero elements can be reduced by a consistent 90% and the overall reduction in number of nonzeros by 60-70%. This results in an improvement of about 40% in terms of number of simplex iterations per unit time. These advantages are well emphasized graphically in Figure 1. The statistics of the problem studied are given in Table 1.

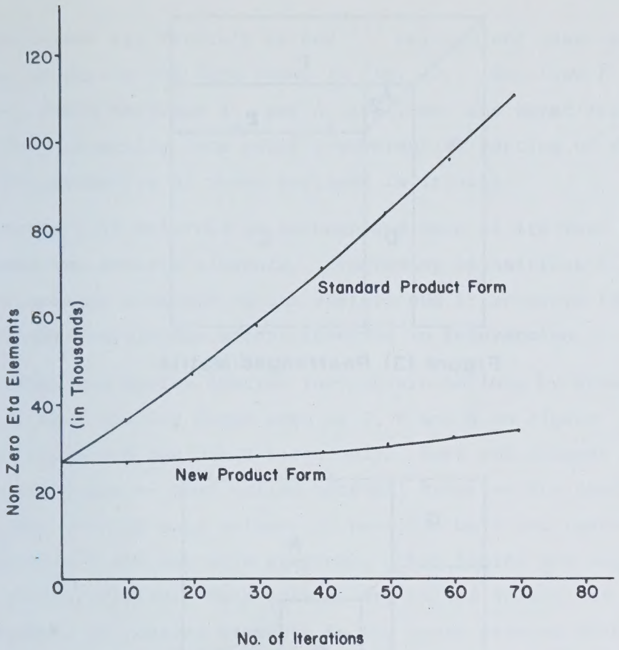
Table 1

PROBLEM STATISTICS

No. of Rows	2,978
No. of Structural Columns	6,333
No. of Elements	47,505
Density	0.25%

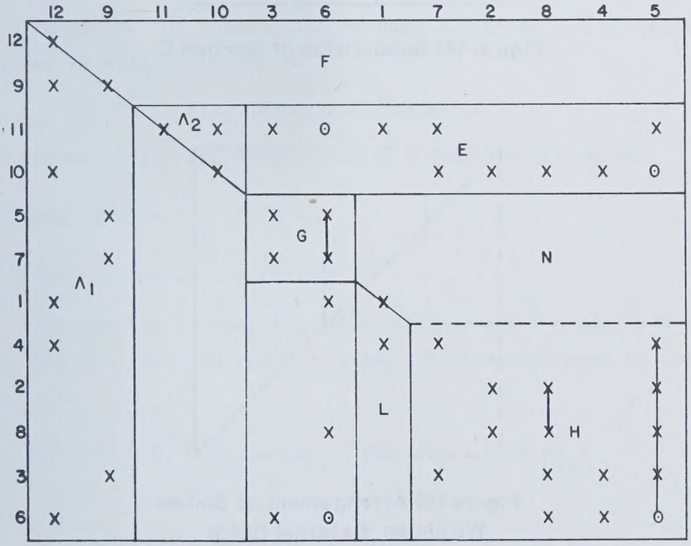
2.2 Hellerman and Rarick's Preassigned Pivot Procedure for Reinversion

The success of reinversion algorthims depends primarily on the fact that even though the basis inverse is unique, its representation in the product form is not. So these algorithms seek a pivot sequence which would give a sparse representation for the basis inverse. The ideal pivot sequence would be one in which the given matrix can be rearranged into a triangular form. However, in a general case, it is almost impossible to obtain this form. Hence, the reinversion algorithms seek an arrangement which is as close to the triangular form as possible.



Grouth of Eta File

FIGURE 1


12 9 11 10 3 6 1 7 2 8 4 5
12
9
11
10
5
7
1
4
2
8
3
6
F
E
G
N
L
H
Λ1
Λ2

Structure after Rearrangement of Columns and Rows

FIGURE 2

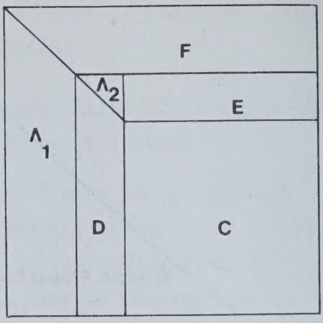


Figure (3) Rearranged Matrix

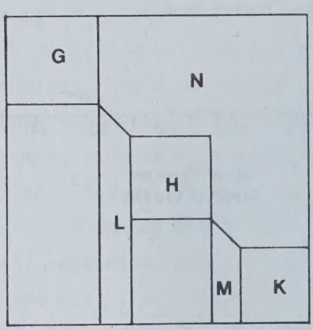


Figure (4) Subdivision of Section C

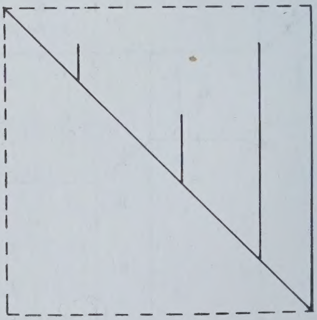


Figure (5) Arrangement of Spikes Within an External Bump

In Hellerman and Rarick's method[12] columns and rows are first rearranged to obtain the form shown in Fig. (3). Sections F and D have only zeros, while sections Λ_1, and Λ_2 are lower and upper triangular in nature. They generally form quite a substantial portion of the whole matrix. Factorization of these sections is trivial.

Submatrix C is referred to as bump and each of its rows and columns has at least two nonzero elements. Processing of sections C and E leads to buildup of nonzero elements in eta vectors and it accounts for the major portion of the computation effort involved in reinversion.

Hellerman and Rarick control the nonzero buildup by breaking Section C into two or more smaller bumps such as G, H and K in Figure (4). It may be noted that Section N again has zeros only. Rows and columns of each of these smaller bumps -- also called external bumps -- are then arranged in such a manner that as many columns as possible have only zeros in the space between section E and the main diagonal. This limits the columns in which nonzeros can be created. Such columns are called spikes and are recognized by the presence of nonzero elements in the space between section E and the main diagonal. The right hand edge of each of the external bumps is always a spike with a nonzero in the uppermost position as shown in Figure (5). The other spikes within an external bump are arranged to be as short as possible, so that they are transformed with the smallest subset of eta vectors. Procedure for subdivision of bump and selection of spikes is very complex, and is well explained in [12].

III. PROBLEM FORMULATION

The economic dispatch problem can be summarily stated as:

$$\text{min. } f(x) = \sum_{i=1}^{NG} f_i(P_i^G)$$

subject to the following constraints:

a) The functional equality constraints - which arise because at each bus the real and reactive power requirements must be satisfied, i.e.

$$P_i^G - P_i^D = P_i^N = V_i \sum_{j=1}^{n} Y_{ij} V_j \text{ Cos } (\delta_i - \delta_j - \psi_{ij}) \quad (3.1)$$

$$Q_i^G - Q_i^D = Q_i^N = V_i \sum_{j=1}^{n} Y_{ij} V_j \text{ Sin } (\delta_i - \delta_j - \psi_{ij}) \quad (3.2)$$

b) The inequality constraints - which arise because of the machine rating and service quality requirements, i.e.

$$P^G_{i_{min}} \leq P^G_i \leq P^G_{i_{max}}$$

$$Q^G_{i_{min}} \leq Q^G_i \leq Q^G_{i_{max}} \qquad (3.3)$$

$$V_{i_{min}} \leq V_i \leq V_{i_{max}}$$

Where, P and Q denote active and reactive powers, V and δ denote voltage magnitude and its angle, Y and ψ denote admittance magnitude and its angle. Further, the superscripts G, D and N are used to signify generation, demand and net injection, respectively, of a quantity at a particular node. Finally, NG represents the number of generating stations and n respresents the number of nodes in the system.

Wolfe's quadratic programming approach was used for solving the economic dispatch problem. For this purpose, the objective function is expressed as:

$$f(x) = \sum_{i=1}^{NG} (C_{o_i} + C_{1_i} P^G_i + \frac{1}{2} C_{ii} (P^G_i)^2)$$

and the equality constraints (3.1) and (3.2) are linearized. This can be accomplished by using Taylor's series expansion and considering only first order terms. The general form of the linearized functional equality constraints then becomes:

$$P^G_i - P^D_i = P^N_i = P^N_{i_o} + \left(\frac{\partial P^N_i}{\partial V}\right)'_o (V-Vo) + \left(\frac{\partial P^N_i}{\partial \delta}\right)'_o (\delta - \delta o) \qquad (3.4)$$

$$Q^G_i - Q^D_i = Q^N_i = Q^N_{i_o} + \left(\frac{\partial Q^N_i}{\partial V}\right)'_o (V-Vo) + \left(\frac{\partial Q^N_i}{\partial \delta}\right)'_o (\delta - \delta o) \qquad (3.5)$$

Where, the subscript o indicates that the quantity is evaluated at the operating point about which the constraint is linearized.

IV. RESULTS

The computer code developed, during the course of this research was tested for its effectiveness in reducing the storage requirements of the basis inverse. It was used to solve the optimal active and reactive power dispatch problem, taking the IEEE standard 14, 30, 57 and 118 bus systems as test cases. The cost function data were taken from [2]. The optimization results, obtained, were very close to the one reported in [2]. However in the case of the new code, a considerable reduction in the storage requirements, of the basis inverse, was achieved. For comparison purposes, the number of nonzero entries in the basis inverse are given in Table 2. The listed values clearly show that the new code becomes more and more effective, in limiting the storage requirements, as the problem size increases. Improvement in the case of the 118 bus system is rather dramatic. The nonzero inverse entries, in this case, are only about 25% of those in Reid's Case. It may further be noted that even for the 118 bus system, the maximum number of simplex iterations, needed in an optimization cycle, are only about 15. These are too small to exploit the full potential of Forrest and Tomlin's updating scheme. So, still better results can be expected when this code is used for solving larger size problems.

Table 2

COMPARISON OF NUMBER OF NONZERO ENTRIES IN THE BASIS INVERSE

No. of Buses	No. of Nonzero Entries in the Basis Inverse	
	Reid's Case	Present Case
14	1,402	833
30	3,148	1,893
57	10,176	3,831
118	36,174	7,669

V. CONCLUSIONS

The special techniques implemented in the new code are extremely effective in reducing the storage requirements of the basis inverse. This will also reduce the computation time appreciably. Hence, there is a great potential for using linear and quadratic programs for solving economic dispatch and system expansion planning problems.

REFERENCES

1. Dayal, G., Grigsby, L.L., Hasdorff, L., "Quadratic Programming for Optimal Active and Reactive Power Dispatch using Special Techniques for Reducing Storage Requirements," Paper A 76 388-9, IEEE PES Summer Meeting, Portland, OR, July 18-23, 1976.

2. Reid, G.F., Hasdorff, L., "Economic Dispatch Using Quadratic Programming," IEEE Trans. PAS-92, pp. 2015-2023, Nov./Dec. 1973.

3. Nicholson, H. and Sterling, M.J.H., "Optimum Dispatch of Active and Reactive Generation by Quadratic Programming," IEEE Trans. PAS-92, pp. 644-654, Mar./Apr. 1973.

4. Nabona, N. and Freris, L.L., "Optimization of Economic Dispatch Through Quadratic and Linear Programming," Proc. Inst. Elec. Eng., Vol. 120, pp. 574-579, May 1973.

5. Wollenberg, B.F., Stadlin, W.O., "A Real Time Optimizer for Security Disptach, "IEEE Trans. PAS-93, pp. 1640-1649, Sept./Oct. 1974.

6. Wells, D.W., "Method of Economic Secure Loading of a Power System," Proc. Inst. Elec. Eng., Vol. 115, No. 8, Aug. 1968, pp. 1190-1194.

7. Knight, U.G.W., "The Logical Design of Electrical Networks using Linear Programming Methods," Proc. Inst. Elec. Eng., PT.A, Vol. 33, 1960, pp. 306-314.

8. Adams, R.N., Laughton, M.A., "Optimal Planning of Power Networks using Mixed Integer Programming," Proc. Inst. Elec. Eng., Vol. 121, No. 2, Feb. 1974, pp. 139-147.

9. Day, J.T., "Forecasting Minimum Production Costs with Linear Programming," IEEE Trans. Power Apparatus and Systems, Vol. PAS-90, No. 2, pp. 814-823, March/April 1971.

10. Day, J.T., Menge, E.E., "The Role of Direct Optimization in Generation Planning," Proceedings of the American Power Conference, Vol. 36, pp. 884-890, 1974.

11. Forrest, J.J.H., Tomlin, J.A., "Updated Triangular Factors of the Basis to Maintain Sparsity in the Product Form Simplex Method," Mathematical Programming 2 (1972), 263-278.

12. Hellerman, Eli and Rarick, Dennis, "Reinversion with the Pre-assigned Pivot Procedure," Mathematical Programming 1 (1971), 195-216.

13. Wolfe, Philip, "The Composite Simplex Algorithm," SIAM Review, Vol. 7, No. 1, Jan. 1965, pp. 42-54.

14. Wolfe, Philip, "The Simplex Method for Quadratic Programming," Econometrica, Vol. 27, 3 (July 1959), pp. 382-398.

A COMPARATIVE STUDY OF POWER NETWORK ALGORITHMS

FERNANDO L. ALVARADO
The University of Wisconsin, Madison, Wisconsin

ABSTRACT

A rational approach for the comparison of competing algorithms to solve the same power network problem is presented. This approach permits the extrapolation of comparisons between algorithms to larger and/or denser systems, and is useful in providing insight into the ultimate performance of each algorithm. The conclusions of the paper support the experimentally well established fact that nodal formulation of power systems problems coupled with sparse matrix techniques are superior to other classes of algorithms.

INTRODUCTION

Advances in algorithms for the analysis of power networks have brought forth a revolution in the type and size of power networks that can be studied by means of digital computers. Perhaps the greatest strides in this direction were achieved by the nodal formulation of network problems coupled with the introduction of sparsity techniques by Tinney et. al. [1]. In many instances however, there are several competing algorithms in use today to solve the same problem.

This paper presents a methodology for the evaluation of the various computer algorithms used for the analysis of power networks. This methodology permits a meaningful comparison between algorithms prior to their actual implementation. The comparisons are not dependent on the computer system used, compiler efficiency or programmer's skill. Furthermore, extrapolation of results to larger and denser systems is possible. The comparisons are based on the concept of computational complexity, which is the most relevant concept for the comparison of algorithms applicable to large systems. The fundamental ideas and some of the results presented in this paper have been recently reported in the power systems literature [2]. This paper extends these ideas to include a study of the effects of network density and ordering scheme used on the computational complexity of the various fundamental operations with sparse matrices. Algorithms to solve practical power system problems are decomposed into sequences of fundamental operations for which the computational complexity can be estimated. The computational complexity of the overall process is then determined. In addition to determining the computational complexity of an algorithm, this methodology can identify algorithm bottlenecks and suggest places where algorithms need improvement. For example, it would serve no purpose to have every step of an algorithm be of order $n^{1.5}$ if there was a single step, no matter how seemingly trivial, of order n^2.

NOMENCLATURE AND BASIC FORMULAS

All comparisons in this paper are based on counts of essential operations (that is, additions, subtractions, multiplications and divisions). Overhead operations (such as pointer manipulations) can require a nontrivial amount of computer time. However, for most properly programmed algorithms the order of complexity of overhead operations is equal to the order of complexity of some essential operation.

Among the most fundamental building blocks of all power systems algorithms are the various operations associated with the solution of the classical linear equations:

$$Ax = b \tag{1}$$

where A is generally nonsingular, sparse, topology-symmetric and with guaranteed diagonals.

One operation necessary in all sparse matrix algorithms is generally the re-ordering of rows and columns to reduce fill-in. A detailed study of the complexity of the re-ordering step is not undertaken in this paper. However, an understanding of the complexity of many of the sorting operations required for re-ordering is available [3]. It is assumed that all necessary "fills" are added during this re-ordering step. The matrix A can thus be considered to be a "perfect elimination" matrix.

The following "building block" steps associated with equation (1) are considered.

1) Factorization. For unsymmetric matrices, this means the decomposition of A into the product LU, where L is lower triangular and U is unit upper triangular. In addition, the inverse of the diagonals of L is stored. Since A is perfect elimination, L and U have the same topology as A and can be stored in place of A. For symmetric matrices, U^tDU factorization is considered, and the inverses of the diagonal matrix D are stored.

2) Repeat solution. For unsymmetric matrices this operation involves two sub-steps: a forward elimination, where the solution to $Ly=b$ is sought, and a back substitution, where the solution to $Ux=y$ is found. For symmetric matrices a similar procedure is necessary.

3) Transposed repeat solution. This involves finding the solution to $A^tx=b$ when the factors of A (not A^t) are already known. It is computationally equivalent to an ordinary repeat solution.

4) Inversion. On occasion, the full inverse of a sparse matrix must be found. This procedure involves one factorization followed by n^2 repeat

solutions. Shortcuts are possible on the forward elimination step, since the b's consist of all zeros and a single one.

5) Sparse inversion. Sometimes only those elements of the inverse of a symmetric matrix that correspond to nonzero elements in A are needed. They can be efficiently found by means of Takahashi, Fagan and Chen's algorithm [4,5].

6) Matrix modification. Frequently, A has been factored and the solution of Ax=b has been obtained. The matrix A is then modified by an amount BRC^t, and the new solution $x + \Delta x$ to:

$$(A + BRC^t)\ (x + \Delta x) = b \tag{2}$$

must be found. If the dimension of R is small the new solution can be found by means of the matrix inversion lemma, which states that:

$$(A + BRC^t)^{-1} = A^{-1} - A^{-1}B(R^{-1} + C^tA^{-1}B)^{-1}\ C^t\ A^{-1} \tag{3}$$

The application of (3) is done using sparsity techniques and without the need of any explicit inverses [2,6,7,8]. The case of greatest interest is when R is a scalar and B and C are vectors. This is the only case analyzed here. In addition, it is assumed that on occasion several modifications from a base case involving the same B vector and different C vectors must be analyzed.

This matrix modification building-block step has a large number of apparently dissimilar practical applications. It can be used when changes occur in the network. It can also be used to make a quasi-symmetric matrix symmetric, and the asymmetry can be taken care of as a "modification" step. Or it can be used to block-diagonalize a matrix, allowing for the independent solution of reduced subnetworks, with a matrix modification step taking care of the interconnection. This latter application fits in the category of so-called "diakoptic" algorithms [6,7].

All of these six fundamental algorithmic building-block steps are well understood, and the exact number of essential operations required for each of the steps is known. Table I summarizes these results. The following notation is used (keep in mind that A is considered already ordered and with all fills already added).

n = dimension of A

r_i = nonzeros in strictly upper triangular portion of the ith row

$\tau = \sum_{i=1}^{n} r_i$ = total number of upper (or lower) triangular nonzero elements

$$\alpha = \sum_{i=1}^{n} r_i^2$$

$$\beta = \sum_{i=1}^{n} i\, r_i$$

COMPLEXITY OF FUNDAMENTAL STEPS

The computational complexity of the fundamental steps 1) – 6) can be determined in terms of the computational complexity of n, τ, α and β. Consider a class of matrices that satisfy the following assumption:

Assumption 1: The maximum number of upper triangular nonzero elements, r_i, in any row is bound by kn^γ, where $0 \le \gamma \le 1$.

For the class of full matrices, $k=1$ and $\gamma=1$. For a class of banded matrices of bandwidth b, $k=(b-1)$, $\gamma=0$.

Based on assumption 1, the following properties about τ, α and β can be established.

Theorem 1: τ is of order no greater than $n^{1+\gamma}$.

Proof:

$$\tau = \sum_{i=1}^{n} r_i$$

$$\tau \le \sum_{i=1}^{n} r_i^{max}$$

$$\tau \le n(kn^\gamma)$$

$$\tau \le kn^{1+\gamma} \qquad \text{q.e.d.} \qquad (4)$$

Theorem 2: α is of order no greater than $n^{1+2\gamma}$.

Proof:

$$\alpha = \sum_{i=1}^{n} r_i^2$$

$$\alpha \le \sum_{i=1}^{n} (r_i^{max})^2$$

$$\alpha \le kn^{1+2\gamma} \qquad \text{q.e.d.}$$

Theorem 3: β is of order no greater than $n^{2+\gamma}$.

Proof:

$$\beta = \sum_{i=1}^{n} i\, r_i$$

$$\beta < \sum_{i=1}^{n} n\, r_i^{max}$$

$$\beta < kn^{2+\gamma} \qquad \text{q.e.d.}$$

Under additional assumptions it can be guaranteed that the above orders for τ, α and β are not merely upper bounds but are also lower bounds as well [2,9].

The expected orders of computational complexity for each of the six building block steps is also summarized in Table I.

DETERMINATION OF γ

Before predictions about the behavior of an algorithm on a certain class of problems can be determined, the value of γ must be estimated. This has been done by selecting several representative matrices among the class of matrices under consideration, ordering them, adding all fills, and using a logarithmic least square error fit to find the values of k and γ from (4) that best match the actual values of τ. Figure 1 illustrates one such fit.

The classes of matrices considered are:

(a) Randomly generated topology-symmetric random matrices, where the probability of finding a nonzero in any position is given and constant.

(b) Randomly generated networks, built by constructing a random tree and adding links randomly. The distribution of the branches per node closely matches the distribution of branches per node in actual power systems.

(c) Actual power networks. The topology matrices of a few actual power networks have also been used.

Representative values of k and γ for each one of these three classes of matrices are given in Table II.

For (a) and (b) it can be expected that the resultant γ be a function of the specified matrix density. The matrix density d is defined as the average value of off-diagonal nonzeros per row in the matrix prior to ordering and addition of fills.

It can also be expected that the type of ordering scheme used to reduce fill in might also influence γ. The three ordering schemes used are schemes 1, 2 and 3 all from [1]. In addition, no ordering at all was studied. The effect of both density and ordering scheme on γ is illustrated in Figure 2. In addition, a curve based on Duff's theoretical formulas [14] for the case of no ordering is also illustrated; excellent agreement with experimental values can be observed.

The conclusion of this section is that for actual networks of practical density and using one of the normal ordering schemes, values of γ in the range from .1 to .25 appear reasonable. A value of .2 is therefore arbitrarily chosen as representative.

APPLICATION TO PRACTICAL ALGORITHMS

Two of the "classical" problems of power systems engineering are power flow and short circuit analysis.

A. The Power Flow Problem

This problem involves the analysis of an electric network with certain nonlinear elements. Early algorithms were based on Gauss method, which soon proved to take an excessive number of iterations in large systems. Another approach, Newton's method, proved successful once adequate techniques for the exploitation of sparsity were developed [1,11]. The main steps in Newton's method are: the construction of a Jacobian matrix, its factorization and a repeat solution at every iteration. The total number of iterations is fairly constant (about 4 or 5) regardless of system size. Clearly, then, based on results from the previous sections the crucial step from the point of view of computational complexity is the factorization step. The computational order for the overall algorithm can thus be expected to be $n^{1+2\gamma}$, with γ between .1 and .25. As seen earlier, the exact number is very much influenced by network density and other properties of the class of networks under consideration.

An evolution of Newton's method is the so-called "fast decoupled" method [12], where a single approximate Jacobian is calculated, factored and used for all subsequent iterations (about 6 or 7, regardless of system size). From the point of view of computational complexity, the crucial step is still the factorization step. However, since it is required only once, it can be expected that its performance relative to Newton's method will get progressively better for large systems. Furthermore, if network modifications are required (as in the case of contingency studies), the new problem does not require the re-factorization of the Jacobian. The first iteration of the first contingency requires an additional repeat solution (a process of order $n^{1+\gamma}$) followed by several operations of order n (depending on how many actual matrix changes are required to analyze the given contingency). Further iterations require merely operations of order n in addition to the ordinary repeat solution of order $n^{1+\gamma}$. Based on the analysis presented on this paper it can then be concluded that, whenever applicable, the fast decoupled method should be preferred both for base case and contingency studies. Contingency studies should be done by matrix modification rather than re-factorization, provided a reasonably small number of matrix modifications are involved in the contingency. These conclusions may, of course, not be applicable to small systems.

B. The Short Circuit Problem

This problem involves the analysis of a slightly different linear electric network for each postulated short circuit condition in the system. It has been shown that this problem can be solved by constructing the nodal (bus) admittance matrix and then finding elements of its inverse. Elements of the inverse correspond to nodal voltages during the various short-circuit conditions. Clearly, then, the entire inverse is needed only if all the currents throughout the entire network are desired for a short circuit postulated at every node in the network. It is unlikely that any useful insight could be gained by such a massive gathering of information. It is more likely that either (1) a detailed study of short circuit currents throughout a large portion of the network is needed for a selected few short circuit locations; or (2) a study of currents through and near the short circuit is needed for a short circuit anywhere in the network. Some early "Z-bus" short circuit algorithms were based on either inversion of "Y-bus" (a process of order $n^{2+\gamma}$ if sparsity techniques are used, n^3 if they are not).[*] Other early algorithms (known collectively as building algorithms) were based on construction of Z-bus directly by repeated applications of the matrix modification formula [3] to a partially built Z-bus. The order of complexity of all the modifications made due to radial type additions is a process of order n^2. The order of complexity of all the modifications made due to loop changes is of order $(b-n)n^2$, where b is the number of branches. Since, generally speaking, a fair assumption about b is that $b=Kn$, then clearly this process is of order n^3. The class of "building algorithms" as known today, therefore, proves to be extremely inefficient.

The preferred procedure for solving sub-problem (1), detailed short circuits at selected fault locations, is merely factorization of Y followed by as amny repeat solutions as needed. If only a few repeat solutions are involved, the order of complexity of this process is $n^{1+2\gamma}$. The preferred procedure for solving sub-problem (2), finding selected currents for all possible fault locations, is sparse inversion, followed, if necessary, by matrix modification to account for a selectee few asymmetric elements. The order of complexity of this process is also $n^{1+2\gamma}$. In either case, modifications in the network can be very efficiently handled by matrix modification, a process of order $n^{1+\gamma}$ once the original desired elements of Z have been found.

---

[*] There is a new class of algorithms for factoring and inverting full matrices that are of order $n^{2.81}$ and are not based on Gaussian elimination [15].

CONCLUSIONS

The main finding of this paper can be summarized as follows. Under the assumption that $\max(r_i)$ is of order n^γ, the value of τ, the number of upper triangular nonzeros in a matrix, is shown to be of order $n^{1+\gamma}$. Similarly, α is shown to be equal to $n^{1+2\gamma}$. The value of γ is experimentally determined to be somewhere around .2 for practical systems. This value of γ is shown to be influenced by network density and by the ordering scheme used.

The practical implications of these results on power systems algorithms for power flow and short circuit analysis is discussed. The importance of matrix modification algorithms that avoid re-factorization of an already factored matrix A is demonstrated.

REFERENCES

[1] W. F. Tinney and J. W. Walker, "Direct Solutions of Sparse Network Equations by Optimally Ordered Triangular Factorization, Proc. of the IEEE 55 , no. 11, November 1967, pp. 1801-1809.

[2] F. L. Alvarado, "Computational Complexity in Power Systems", IEEE Trans. on Power Apparatus and Systems PAS-95, no. 4, July/August 1976, pp. 1028-1037.

[3] D. E. Knuth, "Sorting and Searching", Addison, Wesley, Reading, Mass., 1973.

[4] K. Takahashi, J. Fagan and M. Chen, "Formation of the Bus Impedance Matrix and its Application in Short Circuit Studies", 1973 PICA Conf. Proceedings, Minneapolis, Minn., June 1973, pp. 63-71.

[5] A. M. Erisman and W. F. Tinney, "On Computing Certain Elements of the Inverse of a Sparse Matrix", Comm. of the ACM 18, April 1975, pp. 177-179.

[6] F.L. Alvarado, D.K. Reitan and M. Bahari-Kashani, "Sparsity in Dialoptic Algorithms", IEEE PES Summer Meeting, Portland, Oregon, July 18-23, 1976.

[7] F.F. Wu, "Diakoptics and Triangular Factorization", IEEE PES Winter Meeting, New York, January 27-31, 1975.

[8] W. F. Tinney, "Compensation Methods for Network Solutions by Optimally Ordered Triangular Factorization", IEEE Trans. on Power Apparatus and Systems PAS-91, January/February 1972, pp. 123-127.

[9] F. L. Alvarado, "Computational Complexity of Operations Involving Perfect Elimination Sparse Matrices, submitted for publication.

[10] G. W. Stagg and A. H. El-Abiad, "Computer Methods in Power System Analysis", McGraw-Hill, 1968.

[11] W. F. Tinney and C. E. Hart, "Power Flow Solution by Newton's Method", IEEE Trans. on Power Apparatus and Systems PAS-86, no. 11, November 1967, pp. 1449-1460.

[12] B. Stott and O. Alsac, "Fast Decoupled Load Flow", IEEE Trans. on Power Apparatus and Systems PAS-93, May/June 1974, pp. 859-869.

[13] K. C. Kruempel, "A Study of the Building Algorithm for the Bus Impedance Matrix", PhD Thesis, University of Wisconsin-Madison, 1970.

[14] I. S. Duff, "On the Number of Nonzeros Added when Gaussian Elimination is performed on Sparse Random Matrices", Mathematics of Computation 28, no. 125, January 1974, pp. 219-230.

[15] V. Strassen, "Gaussian Elimination is not Optimal", Numerical Mathematics 13, 1969, pp. 354-356.

TABLE I. OPERATIONS FOR BUILDING BLOCK STEPS

Step	Divisions	Multiplications	Additions/ Subtractions	Order
Factorization (storing inverse of diagonals)				
Symmetric	n	$(\alpha + 3\tau)/2$	$(\alpha + \tau)/2$	α
Topology Symmetric	n	$\alpha + \tau$	α	α
Repeat Solution Transposed R.S.	0	$n + 2\tau$	2τ	τ
Full Inversion				
Symmetric	0 + F.	$2\beta + n(n-1)/2$ + F.	2β + F.	β
Topology Symmetric	0 + F.	$2n\tau + \beta + n(n-1)/2$ + F.	$n\tau + \beta$ + F.	β
Sparse Inversion	0 + F.	$\alpha + \tau$ + F.	α + F.	α
Modification (calculation of Δx)				
First Vector	1 + R.S.	$3n$ + R.S.	$2n - 1$ + R.S.	τ
Subsequent Vectors	1	$2n$	$n - 1$	n

TABLE II. REPRESENTATIVE VALUES OF k AND γ

Type of Matrix	Average Density	Ordering Scheme	k	γ
Random	1.2	2	.940	.196
	1.5	2	.738	.350
Organized random	1.2	2	1.203	.118
	1.5	2	0.872	.289
Actual power systems	1.2	2	1.192	.091

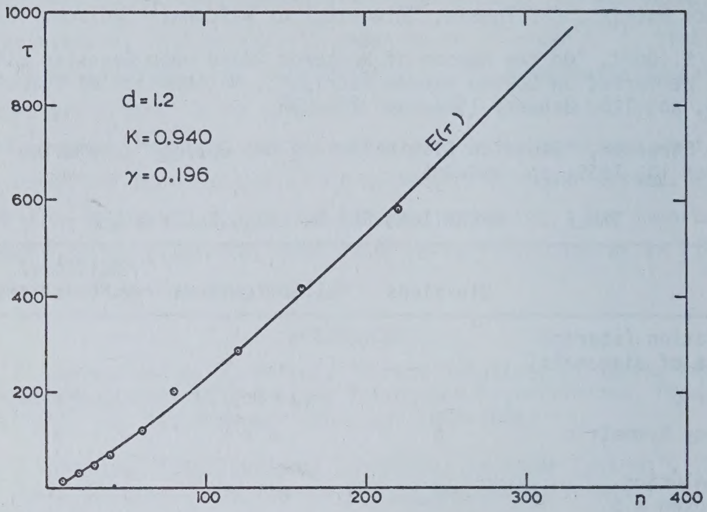


Figure 1. Representative growth of nonzeros in a factored ordered perfect elimination matrix showing curve fit.

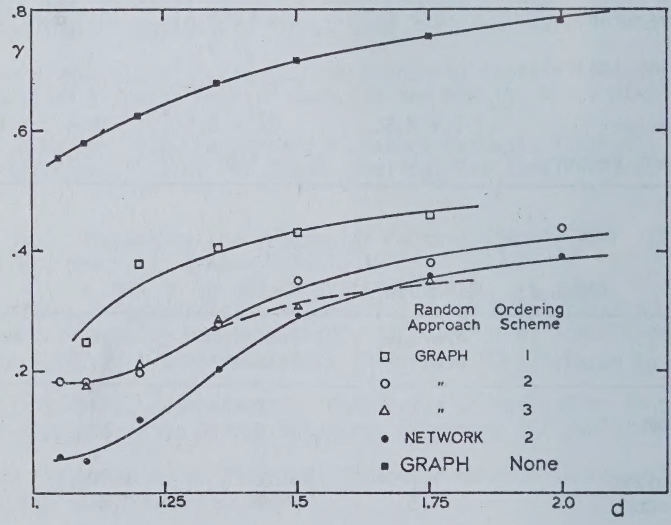


Figure 2. Variation of γ versus density for various ordering schemes and matrix classes.

SECURITY, CONTROL, AND ECONOMIC FUNCTIONS IN POWER SYSTEM CONTROL CENTERS

MARK ENNS
The University of Michigan
Ann Arbor, Michigan

ABSTRACT

This paper summarizes the advanced security, control, and economic functions that have been implemented, are being planned, or are desired for power system control centers throughout the world. Each of the functions is described in terms of its history, mathematical formulation, solution methods, status as of this time, and open problems and needed developments. The functions discussed are: automatic circuit restoration, automatic generation control, automatic voltage control, economic dispatch, interchange scheduling, on-line load flow, optimum power flow, on-line short circuit, security analysis, state estimation, and unit commitment.

INTRODUCTION

Possibly the most interesting development associated with power systems over the past several years is the "new generation" of power system control centers. By general agreement, these start with the Michigan Center which went into operation in June 1969, and now includes approximately 40 centers throughout the world that are either in service or are scheduled to be by the end of 1977. Many others have been contracted for or are in some stage of planning.

The centers vary widely in their size and cost, the number and complexity of the functions performed, the power of the computers employed, number of CRT displays, etc. These differences reflect differences in the size of the systems or pool being controlled, the time of planning for the center (as one might expect, the more recent centers tend to be more elaborate than the earlier ones), and the perceived needs and priorities of the utility or pool. Certain equipment and functional characteristics, however, are common to all the control systems and centers, including:

Data Acquisition and Control Subsystem

- Remote terminal units (RTUs) which form the interface to power system instrumentation and control devices;
- Communication channels (digital) between the RTUs and master stations;
- Master stations which communicate with the system control center.

Computer Subsystem

- Two or more real-time "main" computers;
- "Front end" minicomputers to handle data acquisition and, in some instances, process-control type functions;
- Real-time operating system for the management of computer resources.

Man-Machine Interface Subsystem

- Color CRT displays with limited graphics;
- Light pens, key boards, joy sticks, etc. for man-to-machine communications.

Most control centers include a large wall display, often dynamic to some extent, of the more important features of the system. Many other items, from uninterruptible power supplies to a bomb-proof roof may be a part of a control center, but the three subsystems listed above are the only ones with which we shall be concerned in considering the security, control, and economic functions of the control system. Reference 1 contains an excellent account of the design considerations plus a complete listing of all of the centers in service or under construction as of July 1975. The present paper draws on reference 1 for its nomenclature for the advanced functions discussed below, but we shall be less descriptive and deal more with the mathematical formulation of the functions. Thus that important paper is recommended to give a broader picture of this topic.

OVERVIEW OF FUNCTIONS AND CONTROL HIERARCHIES

The functions can be grouped conveniently, if not altogether accurately, into security, control, and economic as implied by the title. To a considerable extent these functions are performed by different centers which are arranged in a hierarchy. Most large systems are organized into at least two levels and, in one case, as many as four. This may be for technical reasons related to size and complexity or for political or organizational reasons. The control centers and functions with which we are mainly concerned in this paper are usually performed at the "pool" level, with a pool comprising two or more member "companies," These companies will already have had control centers (operation centers, dispatch centers) of their own and will want to retain operators and some functions at that level. These arrangements are variously known as pool/company, master/satellite or, within a single company, company/region.

Many divisions of responsibilities are possible and in use, but the following arrangements are typical:

Pool/Master/Company Functions

Economy: Economic dispatch, interchange scheduling, unit commitment.
Control: Automatic generation control (includes monitoring and control of tie line flows).
Security: State estimation, security analysis.
Other advanced applications programs such as on-line load flow and optimum power flow.

Company/Satellite/Region Functions

Security: Direct monitoring of overloaded lines and equipment, equipment availability and operating limits.
Control: Supervisory control, line switching and tagging.

ADVANCED FUNCTIONS

Table 1 gives a list of the advanced applications functions for security, control, and economic optimization that will be discussed in this paper. All of these functions are being performed or are planned for one or more control centers, although no one center includes them all.

Table 1
Advanced Functions in Power System Control Centers

Control	
Automatic generation control	(AGC)
Automatic voltage control	(AVC)
Automatic circuit restoration	(ACR)
Economy	
Economic dispatch	(ED)
Interchange scheduling	(IS)
Unit commitment	(UC)
Optimum power flow	(OPF)
Security	
On-line load flow	(OLF)
On-line short circuit	(OSC)
Security analysis	(SA)
Optimum power flow	(OPF)

Other

State estimation (SE)

Control Functions

Automatic Generation Control (AGC) Automatic generation control, also known as load-frequency control, was one of the first control functions carried out automatically. The techniques were largely worked out in the 1940s or earlier and implemented during the 1950s by analog equipment. AGC acts to maintain system frequency and scheduled tie-line flows through manipulation of "participating" generating units. The outstanding person in the development of load-frequency control has been, without question, Mr. Nathan Cohn. His work has been reported in many papers over a period of 40 years, of which reference 2 is only the latest.

The basic idea of load-frequency control is to maintain a quantity known as the "area control error" (ACE) at zero, where ACE may be expressed as

$$\text{ACE} = (T_1 - T_0) - 10B(F_1 - F_0) \quad ,$$

where B = frequency bias in MW/0.1 Hz,

T_1, T_0 = actual and scheduled area net interchange, MW and

F_1, F_0 = actual and scheduled frequency, Hz.

If all pools hold their own ACE at zero, frequency and all interchange flows will be at their scheduled values. If one area is short of generation and cannot meet its interchange schedule, then other areas will automatically increase their output to bolster frequency at the expense of some deviation in tie-line flows from scheduled values. The value of the frequency bias B determines the extent of the compromise between the two objectives. Various strategies have been employed to control ACE to zero. The most important control problems turn out to be not ones of conventional control theory (time optimal control, deadbeat response, etc.) but of many practical details such as rate limits on generator response, unavailability of units, and provisions to avoid unnecessary governor action. Because of this, and misunderstanding of the role of AGC in support of system frequency, AGC has proved to be a snare to the unwary control engineer new to the subject.

There has been a renewed interest in research for improved methods of performing AGC. At the time of this writing the Electric Power Research Institute (EPRI) and the Energy Research and Development Administration (ERDA) both have outstanding requests for proposals related to this subject [3,4]. The objective as expressed by the ACE criterion, as differentiated from the means of controlling ACE, has been widely accepted for many years. However, it deals only with total exports or imports of an area. Flows over particular lines or even flows between any two areas are not controlled directly, but are a result of generation and load patterns and the free-flow characteristics of ac transmission systems. This sometimes causes problems of circulating flows which merely increase losses, as around Lake Erie and the Western "doughnut." The only direct controls over these flows are phase-shifting transformers and dc transmission links, both rare and expensive. Of course the flows are affected by the overall design of the ac system and its operating conditions. This appears to be a good area for further research as the problem is troublesome in some places and expensive to correct.

Automatic Voltage Control (AVC) In one sense automatic voltage control is in use universally, as all modern generators have automatic voltage regulators, but not in the sense used here of system AVC. The objective is to maintain a satisfactory voltage "profile" throughout the system and to minimize transmission losses. There are currently no AVC installations in the United States or Europe, but some are planned and there has been one operating in Japan for several years.

AVC can be used to manipulate generator voltages (the set points for automatic voltage regulators) to achieve satisfactory voltage levels throughout the system. This entails, among other things, raising the set points during heavy load periods when voltage drops through transmission lines and transformers are high, and lowering these set points during periods of light load when generating units may actually have to absorb excess reactive power generated by transmission lines. Losses are minimized by maintaining the entire system voltage level at the high end of the permissible range and by avoiding unnecessary circulating currents.

AVC can be considered a subproblem of optimum power flow with the correct set points determined in that more comprehensive context. See the section on OPF for a discussion. It is also important to coordinate the operational AVC with system planning if the benefits of optimum planning, especially for capacitor bank installation, are to be realized in practice.

Automatic Circuit Restoration (ACR) The idea of automatic circuit restoration, under control of a computer and as a result of automatic security analysis, is widely accepted as a desirable goal for power system control centers. The great risk inherent in automatic circuit breaker closing will undoubtedly mean very slow acceptance of this function in practice. Only one company has a definite plan to implement this function and it is not as yet operational there. While it is included in this section on control actions, the intelligence for it comes from security analysis.

Economic Functions

Economic Dispatch (ED) Economic dispatch, as the term has been generally understood for about 30 years, is a procedure for allocating the total load among available generation units so as to (approximately) minimize operating (fuel) costs. At first incremental "station busbar" costs were made equal and later refinements were added to account approximately for transmission losses. Initially analog computers and control devices were used to carry out the economic dispatch function, but all recent installations have been digital. ED is carried out at the relatively slow rates of every 2 to 5 minutes, compared with a rate of 1 to 5 s for AGC, and acts through AGC rather than controlling the generation directly.

The conventional ED problem is formulated as the minimization of a summation of generation costs

$$\min f = \sum_{i \varepsilon G} F_i(P_i)$$

subject to the constraint that the generation should match the load plus losses

$$\sum_{i \varepsilon G} P_i = P_D + P_L$$

and constraints on individual outputs

$$P_{i\,\min} \leq P_i \leq P_{i\,\max} \quad .$$

The $F_i(P_i)$ are cost functions for the individual generators and are usually either piecewise linear or polynomials of fourth or fifth order. P_i is the power output of an individual generator, G is the entire set of generators, P_D is the total load demand, and P_L is the system losses.

The solution is obtained by forming the Lagrangian function

$$L = f + \lambda(P_D + P_L - \sum_{i \varepsilon G} P_i)$$

and differentiating with respect to the P_i and setting the result to zero.

$$\frac{\partial L}{\partial P_i} = \frac{dF_i(P_i)}{dP_i} + \lambda\left(\frac{\partial P_L}{\partial P_i} - 1\right) = 0$$

Thus

$$\lambda = \frac{dF_i(P_i)/dP_i}{1-\partial P_L/\partial P_i}$$

This gives the value of the "system lambda" or equal incremental cost for each generator, modified by the transmission loss penalty factor

$$\frac{1}{1-\partial P_L/\partial P_i}$$

P_L is usually assumed to be a quadratic function of the generation powers

$$P_L = B_{00} + \underline{B}_0^t\underline{P} + \frac{1}{2}\underline{P}^t B\underline{P}$$

from which

$$\frac{\partial P_L}{\partial P_i} = B_{0i} + \sum_{j \varepsilon G} B_{ij}P_j$$

These are the so-called "B coefficients" of economic dispatch. Some type of systematic search and tracking algorithm is used to find the correct value of λ at any time and the corresponding values of the P_i.

Interchange Scheduling Interchange scheduling is meant here to include the computation of generation costs under different conditions and negotiations with other utilities for the purchase or sale or power as well as the actual scheduling in time of the interchanges. The kinds of short-range interchanges entered into routinely by control centers for economic reasons are called

Economy A--transactions that will be executed within an hour or two and small enough that unit commitment will not be affected;

Economy B--transactions that are larger or farther in the future so that load forecasting and unit commitment programs must be used.

The system costs are computed with and without various proposed interchanges to determine the savings if any. If both parties find an interchange beneficial then an agreement is concluded for a certain number of MW, starting and ending at specified times.

Unit Commitment (UC) As with so many other of these functions, unit commitment programs were first developed for simulating system operation in expansion planning studies. In brief, the goal is to put generating units on line so as to meet load requirements plus a "spinning reserve" margin, at minimum cost. Since considerable time (several hours) is required to start most thermal units it is necessary to look ahead for a day or more (up to about five days) to determine an optimal commitment schedule.

In principle the commitment problem can be solved optimally as a dynamic program, and versions with a restricted number of state variables have been used. Branch and bound techniques have also been applied. These methods are summarized along with references in reference 5. The determination of operating costs requires the use of the economic dispatch program as a subroutine.

The most important data for unit commitment is an accurate load forecast. Forecasts are usually made for hourly intervals for one to five days in advance and are based on historical load data, growth trends, and weather forecasts. Despite the advanced probabilistic methods used there are so many uncertainties in weather forecasts plus special situations (e.g., industrial strikes, special television programs) that are difficult to account for in a program, that an experienced operator's judgement may produce a more accurate forecast for any given time period. However, a forecasting program produces a consistent output in a form that can be used by the unit commitment program, and its results can be adjusted by operator intervention. The state-of-the-art in load forecasting is summarized in [6].

There are probably greater savings to be had from an optimum unit commitment than from an optimum economic dispatch (i.e., the optimum loading of units already on line or "committed"). This is definitely a topic open for further research. It has received relatively little attention considering the great economic potential.

Optimum Power Flow (OPF) It should be noted that, in Table 1, optimum power flow is listed under both Economy and Security. In fact many of the functions listed, including automatic voltage control, automatic circuit restoration, economic dispatch, economy A of interchange scheduling, and on-line load flow could be solved as special cases of OPF.

At least two installations use the Lagrange multipliers from the optimal power flow solution to compute penalty functions for economic dispatch. This is an important advance because it allows, for the first time, direct consideration of operating constraints as well as obtaining a more exact economic optimum. In the future it may be expected that optimal power flow will be used directly for real-time, constrained, optimal economic dispatch, but no one has done so to date.

The mathematical nature of OPF is more closely related to the other functions listed under Security and a discussion is deferred to that section.

Security Functions

On-Line Load Flow (OLF) Load flow or power flow has long been a principal tool for power system planning. Now it is being used for security analysis and to aid in short-term scheduling. The power flow program is used in two basic modes. The current-time mode which uses the system data base developed by state estimation is discussed further under security analysis. The power flow program is available also to the operators and engineers to study cases of their choosing. These cases will often be based on forecasted data for a future time of interest, derived either from a specific load forecast or from stored load patterns appropriate to the time (season, day of week, time of day) and adjusted for total expected system load. This mode is used to check system operating conditions for planned outages of equipment. The operator can observe the effects of the outages and then make changes such as reallocation of generation until the power flow solution indicates that the operation will be satisfactory. In general, the program and accompanying data base are available to study any kinds of hypothesized conditions. DyLiacco [1] restricts the term OLF to the current-time mode using real-time data, but the program used is the same and the functions are difficult to separate.

The mathematical nature of a power flow analysis, whether on-line or in a study mode, is a set of simultaneous, nonlinear, algebraic equations with the same number of equations as unknowns. Letting $\underline{X}$ be a vector of dependent variables for which we must solve and $\underline{U}$ a vector of independent or control variables, we have

$$\underline{g}(\underline{X},\underline{U}) = 0$$

where the dimension of $\underline{X}$ is the same as that of $\underline{g}$. The usual choices for the variables are

$$\underline{X} = \left\{\begin{array}{l}\text{angle } \Theta \text{ and voltage V on} \\ \text{each load bus, } \Theta \text{ on each} \\ \text{generation bus}\end{array}\right\}$$

$$\underline{U} = \left\{\begin{array}{l}\text{real power P and reactive} \\ \text{power Q on each load bus,} \\ \text{P and V on each generation} \\ \text{bus, } \Theta \text{ and V at slack bus}\end{array}\right\}$$

The slack bus is a special generation bus at which the angle is specified rather than the real power. One angle may be specified arbitarily since only angle differences have any significance. There must be at least one bus at which real power is not specified because the system losses are unknown; specification of P at all buses would overconstrain the system and no solution would be possible. Providing a slack bus as indicated takes care of both problems at once.

Most modern solutions are based on Newton's method in which $\underline{g}$ is expanded in a Taylor's series in $\underline{X}$ as

$$\underline{g}(\underline{X},\underline{U}) = \underline{g}(\underline{X}_0,\underline{U}) + \left(\frac{\partial \underline{g}}{\partial \underline{X}}\right)_{\underline{X}_0,\underline{U}} \Delta\underline{X} + \dots = 0$$

and solved iteratively for $\Delta\underline{X}$ from

$$J\Delta\underline{X} = -\underline{g}(\underline{X}_0,\underline{U}) .$$

J is the Jacobian matrix $(\partial\underline{g}/\partial\underline{X})_{\underline{X}_0,\underline{U}}$. It is a sparse matrix with the same pattern of nonzero entries as the system admittance matrix and the equations are solved iteratively until $\underline{g} = \underline{0}$ by sparsity methods [7]. The $\underline{g}$ functions are the network equations based on Kirchhoff's laws.

On-Line Short Circuit (OSC) Only one control center has implemented the on-line short circuit function but others have plans for it. Short circuit analysis consists of computing the branch currents that would result from various types of faults (involving one, two, or three of the phase conductors, with or without ground) at different points in the system. The main interest for on-line analysis are the short-circuit interrupting duties that might be imposed on circuit breakers if these faults should occur. While sizing of breakers is clearly a planning problem, it is sometimes possible to control the fault duties by opening certain circuits or by shifting generation.

The short-circuit problem is formulated as a set of simultaneous linear equations, using Thevenin equivalent voltages and impedances. The method of symmetrical components [8] is usually used for the analysis of unbalanced faults.

Security Analysis (SA) The function of security analysis is to determine whether the system is secure in the sense of remaining in a satisfactory operating state upon the occurrence of certain "contingencies." These contingencies are changes in loads, generation or system configuration as caused by branch outages. At this time only steady-state SA is attempted by any of the control centers. Transient and dynamic SA may be performed in the future but is beyond the capability of present-day computing systems.

The analysis is performed by some variation of a power flow program. Some centers use precalculated "distribution factors" which, based solely on network impedances, give approximate changes in flows in all system branches for the contingencies considered. More exact solutions may be obtained by the "dc power flow" (a linearized power flow solution) or a full power flow solution. The solutions for branch outages may be obtained more rapidly through use of the Matrix Inversion Lemma [9]. This combined with the fast decoupled load flow [9] gives the most rapid means of getting accurate solutions for a large number of contingencies.

The basic data for SA is, or should be, provided by state estimation. Unfortunately, SE is performed by any one operating unit only on its own system, while information is also required on other interconnected systems for SA. Furthermore, the external system is generally so large that it must be reduced to a smaller "equivalent" system for SA to be practical. This raises a host of problems whose solution is incomplete in regard to the exchange of information among control centers and in the development of on-line equivalents. The problem of information exchange will be addressed in a COPs (Common Operating Problems) session at the 1977 Winter Annual Meeting of the Power Engineering Society, IEEE [10].

The determination of corrective strategies in the case of insecurity may also be considered to be a part of SA. However, this can best be performed by an optimum power flow program, discussed next.

Optimum Power Flow (OPF) Several of the functions discussed above can be accomplished simultaneously through solution of optimal power flows. This is a steady-state optimization of an objective function, usually operating cost, while meeting electrical network constraints imposed by Kirchhoff's laws, equality constraints on system loads, and inequality constraints on voltages, line loadings, etc., applied for security or operational reasons.

It fits naturally with security analysis by finding corrective strategies for insecure cases. It can also, in principle, solve the economic dispatch problem (generator outputs with fixed voltages) and automatic voltage control (generator voltages with fixed power outputs) either separately or simultaneously. Interchange scheduling and unit commitment can use QPF as they now use ED and, in some centers, OLF. Thus a sufficiently fast and reliable OPF could replace several of the individual programs used by most control centers.

OPF is in practice more closely related to the conventional power flow problem than to general optimization problems. That is, the successful computer programs are based on power flow programs [11,12,13] with only modest extensions to deal with the optimization. In essence, the formulation for OPF is the same as for OLF except that generator real powers and voltages are used to optimize the objective function instead of being fixed *a priori*. This is accomplished either by a first-order (gradient) computation [11,12] or a second-order method using a Hessian matrix [13]. In any case, most of the computation is associated with a power flow subproblem which must now be performed several times as new values are found for the $\underline{U}$ (control) variables. This topic is summarized well in references [5] and [14].

The most difficult problem with OPF programs is in the handling of functional inequality constraints on dependent variables. In practical programs they are sometimes handled by penalty functions. Another promising approach, the "generalized reduced gradient" method, solves for the inequality constraints directly by making a change in basis variables when a constraint becomes active. Reference [15] proposes a related method while using a different (complex) formulation of the constraints.

Other Functions

Many other functions could be listed here, but the one of greatest interest as an advanced function is state estimation. It is placed in a section of its own because it does not fit properly in any of the other three categories and, in fact, provides the data needed for nearly all of them.

State Estimation (SE) State estimation is perhaps the most remarkable development to have come from the work of the past eight-odd years on power system operation and control. It involves sophisticated network analysis and computational procedures, but is wholly original to the operations problem and has no role in system planning. In its essence it is a procedure for converting telemetered transmission network measurements and switch positions into a complete and reliable estimate of the network electrical state. The electrical state is conveniently taken as the complete set of system node-to-datum or bus (complex) voltages because all other electrical quantities can be computed from these directly. The measurements are usually real and reactive power flows at the ends of transmission lines and through transformers, bus voltage magnitudes, and bus real and reactive power injections.

The basic model assumed to relate the measurement and state variables is

$$\underline{z} = \underline{f}(\underline{x}) + \underline{v} \quad ,$$

where $\underline{z}$ = vector of measurements,
$\underline{x}$ = vector of state variables (complex bus voltages),
$\underline{f}$ = network equations based on Kirchhoff's laws relating the $\underline{z}$ and $\underline{x}$ variables and
$\underline{v}$ = vector of measurement noise.

A quadratic function of the residual $[\underline{z} - \underline{f}(\underline{x})]$ is formed and values of $\underline{x}$ are computed that minimize it in a least-squares sense with respect to the error covariance matrix. Redundant measurements, usually 100 percent or more, allow identification of "bad data" and their removal from the set of measurements. Thus the state estimator serves as a check on the data acquisition system as well as presenting best estimates of the variables. The estimate forms a real time data base for the power flow and other security analysis and optimization programs.

Several solution schemes have been proposed which differ slightly in their theoretical aspects and more significantly in their computational requirements. Many papers have been written on the subject since the first ones appeared in 1969. Reference 16 contains an excellent summary of the subject. More recent work [17] shows a fast decoupled approach to be most effective computationally.

ACKNOWLEDGEMENT

Much of this paper is based on an earlier paper by the same author presented at the Eighth Annual Princeton Conference in 1974 [18].

REFERENCES

1. T. E. DyLiacco, "System Control Center Design," a report prepared for the United States Energy Research and Development Administration (ERDA), July 1975.
2. N. Cohn, "Techniques for Improving the Control of the Bulk Power Transfers on Interconnected Systems," IEEE Trans. on Power Apparatus and Systems, vol. PAS-90, No. 6, pp. 2409-2419, November/December 1971.
3. Electric Power Research Institute (EPRI) Request for Proposal RFP5313, "Improved Economic Operation of Power Systems," September 10, 1976.
4. Energy Research and Development Administration (ERDA), "Automatic Generation Control--A Systems Approach," Section VI of Systems Engineering for Power," Program Report ERDA 76-82/1.
5. A. M. Sasson and H. M. Merrill, "Some Applications of Optimization Techniques to Power Systems Problems," Proc. IEEE, vol. 62, No. 7, pp. 959-972, July 1974.
6. F. D. Galiana, "Short-Term Load Forecasting," a report prepared for the United States Energy Research and Development Administration (ERDA), July 1975.
7. W. F. Tinney and C. E. Hart, "Power Flow Solution by Newton's Method," IEEE Trans. on Power Apparatus and Systems, vol. PAS-86, No. 11, pp. 1449-1460, November 1967.
8. P. M. Anderson, Analysis of Faulted Power Systems, Ames: Iowa State University Press, 1973.
9. B. Stott and O. Alsac, "Fast Decoupled Load Flow," IEEE Trans. on Power Apparatus and Systems, vol. PAS-93, No. 3, pp. 859-867, May/June 1974.
10. COPs paper on "External Information Requirements for System Operations," to be presented at the 1977 IEEE PES Winter Meeting.
11. H. W. Dommel and W. F. Tinney, "Optimal Power Flow Solutions," IEEE Trans. on Power Apparatus and Systems, vol. PAS-87, No. 10, pp. 1866-1876, October 1968.
12. O. Alsac and B. Stott, "Decoupled Algorithms in Optimal Load Flow Calculations," IEEE paper A 75 545-4.
13. A. M. H. Rashed and D. H. Kelly, "Optimal Load Flow Solution Using LaGrange Multipliers and the Hessian Matrix," IEEE Trans. on Power Apparatus and Systems, vol. PAS-93, No. 5. pp. 1292-1297, Sep./Oct. 1974.
14. H. H. Happ, "Optimal Power Dispatch--A Comprehensive Survey," Paper F 76 440-8, presented at the IEEE PES Summer Meeting, Portland, OR, July 18-23, 1976.
15. W. R. Barcelo, W. W. Lemmon, and H. R. Koen, "Optimization of the Real-Time Dispatch with Constraints for Secure Operation of Bulk Power Systems," Paper F 76 510-8, presented at the IEEE PES Summer Meeting, Portland, OR, July 18-23, 1976.
16. F. C. Schweppe and E. Handschin, "Static State Estimation in Electric Power Systems," Proc. IEEE, vol. 62, No. 7, pp. 972-982, July 1974.
17. H. P. Horisberger, J. C. Richard, and C. Rossier, "A Fast Decoupled Static State-Estimator for Electric Power Systems," IEEE Trans. on Power Apparatus and Systems, vol. PAS-95, No. 1, pp. 208-215, Jan./Feb. 1976.
18. M. Enns. "Applications Program in Control Center Computers," Proc. of the Eighth Annual Princeton Conference on Information Sciences and Systems, Princeton, NJ, pp. 256-260, March 1974.

THE Z-MATRIX IN POWER SYSTEM NETWORK ANALYSIS

H.E. BROWN
North Carolina State University - Raleigh, NC

INTRODUCTION

It was recognized in the early development of computer programs, that the Z-matrix contained much more information about a network, than did the Y-matrix, because it is full whereas the Y-matrix contains a high percentage of zero elements. However, the development of techniques using the Z-matrix did not take place as quickly as Y-matrix methods, because of the difficulty in forming it. Originally the Z-matrix was computed by inversion of the Y-matrix.

The development of the algorithm 1. by which a network is assembled by adding one element at a time, and the Z-matrix modified for each line addition, made the Z-matrix easily available and superior to the Y-matrix for certain applications. Originally these techniques were limited to networks containing less than 250 nodes, because the Z-matrix is full and therefore imposes a severe requirement on the computer memory.

The extention of short circuit studies to large systems (2500 nodes) was achieved by an axis discarding technique. 2. The development of this technique was also crucial to contingency analysis methods that were developed later. The first high speed contingency analysis method used the Z-matrix to test for over-loaded lines, when certain other lines were removed from the network; or when there was a change in the generator inputs to the network. 4. Later a method was developed for predicting low voltages at nodes of the network when lines are removed from the system. 5. The success of this method also depended upon the Z-matrix and the axis discarding technique.

THE Z-MATRIX BUILDING ALGORITHM

In a network one node is chosen as a reference. The driving point impedance is obtained between all other nodes and this reference node. These impedances are the diagonal elements of the Z-matrix. The off-diagonal elements are the transfer impedances of each node with respect to every other node, except the reference. The network is assembled by adding one element at a time. Each impedance, when it is added to the network, must have a continuous path back to the reference. The matrix is modified for each line addition. 1., 3.

Three cases can occur in making the line additions:

1. A radial line from the reference to a new node.
2. A radial line from a node in the network to a new node.
3. A line between two nodes in the network.

In case 1. a new axis is added to the matrix, whose driving point impedance, Z_{kk}, is equal to the impedance of the line being added to the network. The off-diagonal elements Z_{KN} and Z_{NK}, where N = 1, 2, ..., K-1, are all equal to zero.

In case 2 a new axis is added to the matrix. The diagonal element of the new axis, Z_{KK}, is equal to the diagonal element, Z_{NN}, corresponding to the existing node N, plus the impedance of the line being connected, z_{NK}. The off-diagonal elements of the new axis are equal to the corresponding elements of the existing axis.

$$\begin{aligned} Z_{Ki} &= Z_{Ni} \\ Z_{iK} &= Z_{iN} \end{aligned} \qquad i = 1,2,\ \ldots,\ K\text{-}1 \tag{1.}$$

Addition of a line between two nodes already in the system, case 3, is a loop closing line. A loop axis is added to the matrix. The diagonal element of the loop axis for addition of a line between M and N is obtained by equation (2.).

$$Z_{LOOP\text{-}LOOP} = Z_{MM} + Z_{NN} - 2Z_{MN} + z_{LINE\ M=N} \tag{2.}$$

All quantities in equation (2.) are matrix elements except $z_{LINE\ M\text{-}N}$, which is the impedance of the line being added.

The off-diagonal elements of the loop axis are obtained by means of equations (3.) and (4.).

$$Z_{LOOP\text{-}i} = Z_{Mi} - Z_{Ni};\ i = 1,2,\ \ldots,\ LOOP\text{-}1 \tag{3.}$$

$$Z_{i\text{-}LOOP} = Z_{iM} - Z_{iN};\ i = 1,2,\ \ldots,\ LOOP\text{-}1 \tag{4.}$$

All elements of the matrix are modified by a Kron reduction, equation (5.)

$$Z'_{ij} = Z_{ij} - Z_{i\text{-}LOOP}Z_{LOOP\text{-}j}/Z_{LOOP\text{-}LOOP} \tag{5.}$$

After the Kron reduction all elements of the LOOP axis are erased.

The network is assembled one line at a time and the matrix modified after each line addition, to reflect the change in the network configuration.

AXIS DISCARDING

By examining equations (1, 2, 3, and 4) it is seen that the only elements taking an active part in the matrix modification correspond to nodes to which a radial line or a loop closing line is being connected. As soon as all lines have been connected to a node, the elements in the corresponding row and column will never again take an active part in the algebra of the matrix formation. If the node is not required in the network analysis, which will be made after the matrix is completed, the corresponding matrix elements need not be saved. Thus the matrix axis of a node that is not of interest can be erased when the last line has been connected to it. The matrix can be closed up to conserve computer memory.

SHORT CIRCUIT STUDIES

In Short Circuit Studies, the common node behind all generator reactances is selected as the reference node. 1,3. The matrix is formed and axes, that are not of interest, are discarded during the matrix building. Each diagonal element is the Thevenin equivalent impedance between the corresponding node and the reference. In a power system all nodes are normally at a voltage very close to unity. The Thevenin voltage for every node is therefore assumed to be 1.0 pu. The total short circuit current at any node N is approximately the value obtained from equation (6.)

$$I_N = \frac{1.0}{Z_{NN}} \tag{6.}$$

It is generally assumed that the customer load currents are so small, compared to the short circuit current, that they can be ignored. For a particular fault condition the current vector, of matrix equation (7.) has a single non-zero current.

$$ZI = E \tag{7.}$$

The complete analysis of the system is now possible, since for each fault condition, the complete voltage profile is defined. The flow in line P-Q, whose impedance is $z_{lineP-Q}$, for a short circuit at node N is computed by use of equation (8.).

$$I_{P-Q} = \frac{Z_{NQ} - Z_{NP}}{Z_{NN}\, z_{lineP-Q}} \tag{8.}$$

Once the matrix has been formed the complete analysis, total short circuit current at each node and all line flows can be obtained by simple arithmetic calculations using equations (6.) and (8.).

OVER-LOAD CONTINGENCY ANALYSIS

Contingency analysis is the determination of the power flow in the branches of a network under normal loading conditions, when one or more lines have been removed from the network. Contingency analysis also includes the determination of overloaded lines, that results from changes in the generation inputs to the network. In this work it is convenient to make ground the reference and a single arbitrarily chosen node of the system is connected to the reference with any reasonable impedance, e.g. 1.0 + j1.0. 4. The impedance of the connection to the reference is unimportant because its effect is eliminated in the calculation. See equation (16.)

A power flow study, with all lines in normal operating condition is available in the computer library. It is desired to know the increment of load that will be picked up when one or more lines are involved in line changes.

A matrix is formed, and the axes of those nodes corresponding to the termini of lines involved in line changes in the several studies to be made and the termini of lines to be tested for overloading are saved. In a particular case, assume that it is desired to remove lines P-Q and R-S, and to test lines I-J and M-N for over-load condition.

In the short circuit study the single current I_N satisfied a single constraint of equation (9.).

$$E_N = 1.0 \tag{9.}$$

In the contingency case being considered here, we have two conditions to be satisfied, that the voltage differences across the lines P-Q and R-S must be the same voltage differences that exist in the base case load flow. The current vector of equation (8.) must satisfy these two conditions. Two conditions are to be satisfied and therefore two currents can be uniquely determined.

Extract out of the matrix, that has been saved, the matrix of equation (10.) and multiply it by a current vector which is to be determine.

$$\begin{array}{cc} & \begin{array}{cccc}P & Q & R & S\end{array} \\ \begin{array}{c}P\\Q\\R\\S\end{array} & \left[\begin{array}{cccc} & & & \\ & Z & & \\ & & & \\ & & & \end{array}\right]\end{array} \left[\begin{array}{c}I_P\\0\\I_R\\0\end{array}\right] = \left[\begin{array}{c}E_P\\E_Q\\E_R\\E_S\end{array}\right] \qquad (10.)$$

Subtract every second row of the Z-matrix and every second element of the E matrix from the row ahead. Delete every second column of the Z matrix and every second element of the I matrix.

$$\begin{array}{cc} & \begin{array}{cc}P & R\end{array} \\ \begin{array}{c}P-Q\\R-S\end{array} & \left[\begin{array}{c} Z \end{array}\right]\end{array} \left[\begin{array}{c}I_P\\I_R\end{array}\right] = \left[\begin{array}{c}E_P-E_Q\\E_R-E_S\end{array}\right] \qquad (11.)$$

Invert the 2 x 2 matrix and obtain

$$\left[\begin{array}{c}I_P\\I_R\end{array}\right] = \begin{array}{cc} & \begin{array}{cc}P & R\end{array} \\ \begin{array}{c}P-Q\\R-S\end{array} & \left[\begin{array}{c} Z^{-1} \end{array}\right]\end{array} \left[\begin{array}{c}E_P-E_Q\\E_R-E_S\end{array}\right] \qquad (12.)$$

If these elements are used in equation (10.) none of the voltages will have any real significance. These two currents only satisfy the two constraints, that E_P-E_Q and E_R-E_S will have the same values as obtained from the base case load flow.

Extract from the Z-matrix, that has been saved, the matrix of equation (13.)

$$\begin{array}{cc} & \begin{array}{cc}P & R\end{array} \\ \begin{array}{c}I\\J\\M\\N\end{array} & \left[\begin{array}{cc} & \\ & Z \\ & \\ & \end{array}\right]\end{array} \left[\begin{array}{c}I_P\\ \\ I_R\\ \end{array}\right] = \left[\begin{array}{c}E_L\\E_J\\E_M\\E_N\end{array}\right] \qquad (13.)$$

Again subtract every second row of Z from the row before it and every second element of the E matrix from the element before it as shown in equation (14.).

$$\begin{array}{cc} & \begin{array}{cc}P & R\end{array} \\ \begin{array}{c}I-J\\M-N\end{array} & \left[\begin{array}{c} Z \end{array}\right]\end{array} \left[\begin{array}{c}I_P\\I_R\end{array}\right] = \left[\begin{array}{c}E_I-E_J\\E_M-E_N\end{array}\right] \qquad (14.)$$

The original matrix is modified for the line removals using the building algorithm by adding lines with the opposite algebraic sign of the original data. The new network condition is given by equation (15.).

$$\begin{array}{cc} & \begin{array}{cc} P & R \end{array} \\ \begin{array}{c} I\text{-}J \\ \\ M\text{-}N \end{array} & \left[\begin{array}{cc} & \\ \quad Z' & \\ & \end{array}\right] \end{array} \begin{bmatrix} I_P \\ \\ I_R \end{bmatrix} = \begin{bmatrix} E'_I - E'_J \\ \\ E'_M - E'_N \end{bmatrix} \qquad (15.)$$

The change in voltage across line I-J caused by the removal of lines P-Q and R-S is given by equation (16.)

$$\Delta E_{I-J} = (E_I - E_J) - (E'_I - E'_J) \qquad (16.)$$

Although $(E_I - E_J)$ and $(E'_I - E'_J)$ do not have any real significance, the difference ΔE_{I-J} and the increment of current, equation (17.) are accurate to within 1% of the difference, that would actually be produced in the network, as a result of the redistribution of current when the lines are removed.

$$\Delta I_{I-J} = \Delta E_{I-J} / z_{\text{line } I-J} \qquad (17.)$$

The increment of current picked up by line MN is computed in the same manner.

Extract from the matrix, that has been saved, the elements of matrix equation (18.) to investigate the increment of current picked up by lines I-J and M-N, when the generation at node G_1 is increased and the generation at node G_2 is decreased by the same amount.

$$\begin{array}{cc} & \begin{array}{cc} G_1 & G_2 \end{array} \\ \begin{array}{c} I \\ J \\ M \\ N \end{array} & \left[\begin{array}{cc} & \\ \quad Z & \\ & \\ & \end{array}\right] \end{array} \begin{bmatrix} I_{G1} \\ \\ -I_{G1} \\ \\ \end{bmatrix} = \begin{bmatrix} E_I \\ E_J \\ E_M \\ E_N \end{bmatrix} \qquad (18.)$$

Subtract every second row of the Z and E matrices to give equation (19.).

$$\begin{array}{cc} & \begin{array}{cc} G1 & G2 \end{array} \\ \begin{array}{c} I\text{-}J \\ \\ M\text{-}N \end{array} & \left[\begin{array}{cc} & \\ \quad Z & \\ & \end{array}\right] \end{array} \begin{bmatrix} I_{G1} \\ \\ -I_{G1} \end{bmatrix} = \begin{bmatrix} E_L - E_J \\ \\ E_M - E_N \end{bmatrix} \qquad (19.)$$

The current increment is obtained by equation (20.)

$$\Delta I_{I-J} = \frac{E_I - E_J}{z_{\text{line } I-J}} \qquad (20.)$$

LOW VOLTAGE CAUSED BY A CONTINGENCY

Consider the 6-bus load problem of Ward and Hale as shown in Fig. 1.

Generators are connected to nodes 1 and 2 and supply the customer loads connected to the other nodes of the system. With a fixed voltage magnitude and angle at one node (e.g. node 1), a solution has been reached when the voltage of the other nodes has been determined, such that, Kirchhoff's law is satisfied for every node.

A Z-matrix formulation of the load flow problem uses ground as the reference. Customer loads are represented by impedance to ground as shown in Fig. 2. The load impedances are computed by use of equation (21.) in which E is assigned a reasonable value, for example 1.0 pu.

$$Z = \frac{E^2}{P-jQ} \tag{21.}$$

The node voltages take on values different than the assumed value used in equation (21.) during the iterative solution. The load impedances will then not consume the prescribed values of power. The differences between the desired value of load and the load in the ground tie is made equal to zero by a direct injection of current into the node. See equation (22.).

$$P + jQ = \frac{E^2}{Z^*} + EI^* \tag{22.}$$

Note * indicates complex conjugate.

Thus when solution has been reached the current vector is full. Here N currents determine N voltages.

In the method being described here, 4 it is desired to know if one node (for example node 5) has low voltage when one or more lines have been removed from the network. Node 1 is a generator and has a constant voltage magnitude. The matrix of the network is formed with a single, arbitrary ground tie impedance (1.0 + j1.0) connected to the node to be tested for low voltage. A single generator is connected to the constant voltage node. See Fig. 3.

The arbitrary ground tie must be replaced by an equivalent ground tie as it appears looking in at node 5. An equivalent generator must be determined for node 1. The reduce network is as shown in Fig. 4.

We must now determine I, (the equivalent generator) and the equivalent load impedance, Z, as seen by node 5 such that two conditions E_1 and E_5 of the base case load flow are satisfied. See equations (23.) and (24.).

$$I_1Z_{11} = I_1(Z_{11} - Z_{55}) + I_1Z = E_1 \tag{23.}$$

$$I_1Z = E_5 \tag{24.}$$

Substitution of (24.) into (23.) gives equation (25.)

$$I_1 = \frac{E_1 - E_5}{Z_{11} - Z_{55}} \tag{25.}$$

Using this value of I_1 in equation (24.) gives the value for Z. Equation (26.).

$$Z = \frac{E_5}{I_1} \qquad (26.)$$

The Z-matrix of the network is modified for the removal of one or more lines. It is argued, that removing lines does not alter the ground impedance, Z, as seen by node 5. (If the lines being removed have considerable line charging the line charging ground ties must also be removed at the time the lines are removed and the matrix modified accordingly.)

A new injection current I_1' must be determined which will satisfy equation (27.).

$$I_1' = E_1/(Z_{11}' - Z_{55}' + Z) \qquad (27.)$$

The magnitude of the fixed voltage node is constant and the equivalent ground tie impedance is known. The voltage on node 5 after the lines have been removed is given by equation (28.).

$$E_5' = I_1'Z \qquad (28.)$$

To test the voltage reduction on another node (for example node 6), an arbitrary ground tie is made to node 6; the ground tie is removed from node 5 and the simple process repeated.

This method is especially important in systems with weak transmission systems. It was tested extensively on the Furnas system of Brazil with excellent results. This method predicted 8.5% voltage drops while the Newton-Raphson method indicated 10.0%. Which answer is correct on the actual system is not known. The indication is, that in these cases there is real system trouble, and the results are obtained at 100 times the speed of full load flow solutions.

REFERENCES

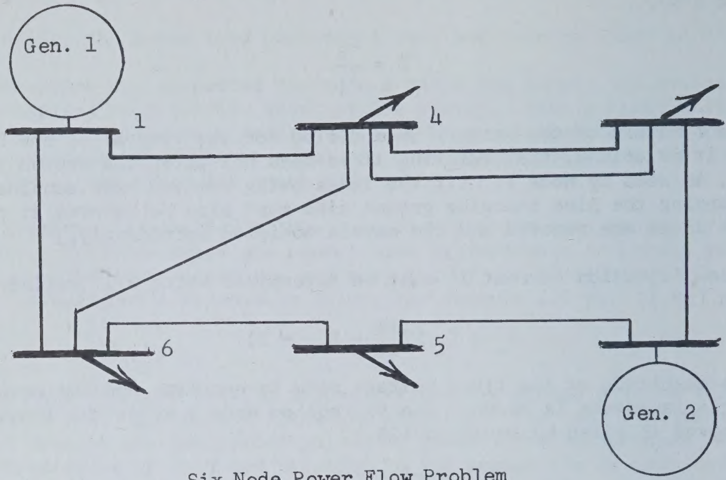
1. H. E. Brown, C. E. Person, L. K. Kirchmayer, and G. W. Stagg, Digital Calculations of Three Phase Short Circuits by Matrix Method, Trans. AIEE, Vol. 79 Part III (1960) pp 1277-81.

2. H. E. Brown, and C. E. Person, Short Circuit Studies of Large Systems by the Impedance Matrix Method, IEEE, PICA Conf. Proc. (1967). Pittsburgh, PA pp 335-42.

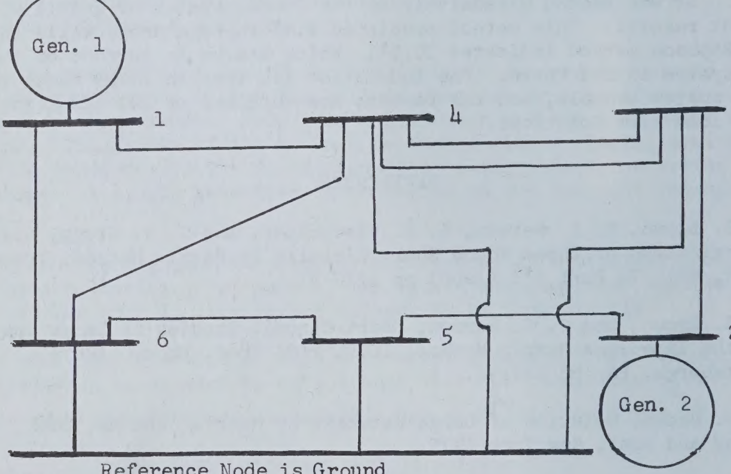
3. H. E. Brown, Solution of Large Networks by Matrix Methods, Book, John Wiley and Sons, New York 1975.

4. H. E. Brown, Interchange Capability and Contingency Evaluation by a Z-matrix Method, IEEE, Trans. PA&S (Sept, Oct 1972) page 1927.

5. H. E. Brown, Low Voltage Caused by Contingencies Predicted by a High Speed Z-matrix Method, IEEE Summer Power Meeting, 1976, Portland. Paper F76-374 to be published later in IEEE, Trans. PA&S.

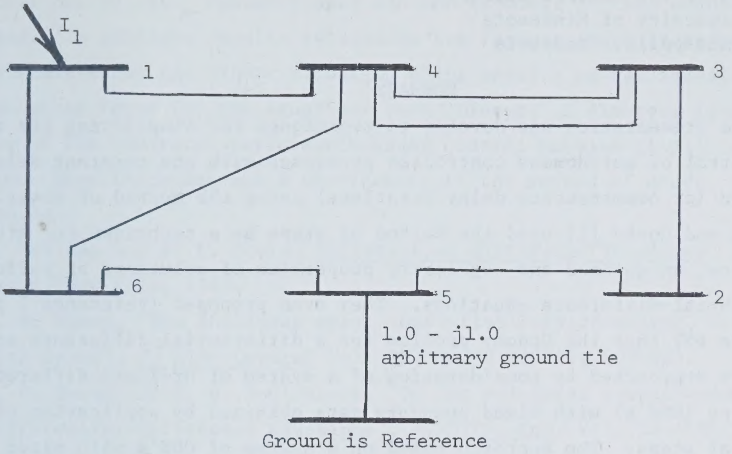

Gen. 1
1
4
3
6
5
2
Gen. 2

Six Node Power Flow Problem
Fig. 1

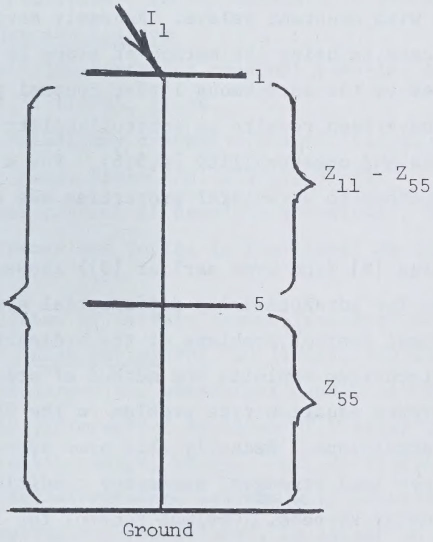

Gen. 1
1
4
3
6
5
2
Gen. 2

Reference Node is Ground
Z-Matrix representation showing ground ties

Fig. 2


I_1
1
4
3
6
5
2
1.0 - j1.0
arbitrary ground tie
Ground is Reference

Contingency Analysis Representation of 6 Node Network
Fig. 3


I_1
1
$Z_{11} - Z_{55}$
Z_{11}
5
Z_{55}
Ground

Network Used in the Analysis
Fig. 4

THE METHOD OF STEPS IN THE THEORY OF CONTROL OF TIME DELAY PROCESSES*

E. B. LEE
Center for Control Sciences
University of Minnesota
Minneapolis, Minnesota

SUMMARY

The presentation was devoted to techniques for simplifying the theory for control of autonomous controlled processes with one constant delay duration (or commensurate delay durations) using the method of steps. Bellman and Cooke [1] used the method of steps as a technique for studying existence, uniqueness and regularity properties of solutions of various differential-difference equations. They even proposed (reference 1 page 95 exercise 65) that the Cauchy problem for a differential-difference equation could be approached by consideration of a system of ordinary differential equations (ODE's) with mixed boundary data obtained by application of the method of steps. The approach based on a system of ODE's with mixed boundary data has been successfully used by Zmood [2,3] in studying controllability and pointwise completeness properties of the autonomous linear control process with constant delays. Recently several researchers have had similar success in using the method of steps to answer other structural questions related to the autonomous linear control process with constant delays. There have been results on controllability with piecewise constant controlled inputs and observability [4,5,6]. For a detailed account of the various contributions to structural properties see the lecture notes of Manitius [7].

In 1968 Warga [8] (and even earlier [9]) showed how certain optimal control problems for advanced-delay differential systems could be reduced to standard optimal control problems of the ordinary differential equation type. Warga's technique exploits the method of steps to reduce the differential-difference equation type problem to the ODE type problem with mixed boundary conditions. Recently this same approach was used by Olbrot [10] to obtain new (and stronger) necessary conditions for the optimal control of the delay systems. Optimal control for linear autonomous control processes with constant delays and the quadratic performance index has been treated by Korytowski [11] using the method of steps. He is able to obtain a simple analytic formula for the optimal operator in the feedback controller synthesis.

*Research supported by the Air Force under Grant Number AF-AFOSR-72-2243.

The presentation provided an overview of the method of steps in the simplification of the theory of control of the autonomous control process with constant delays. Disadvantages and shortcomings of this method were included. In addition results related to the further simplifications that can be achieved in the linear situation using results on realizability and canonical forms for the equations were offered. A discrete time version of the quadratic performance index control problem displayed a further simplification and a shortcoming of the method of steps.

References

1. R. Bellman and K. L. Cooke, Differential-Difference Equations, Academic Press, New York, 1963.
2. R. B. Zmood, "The Euclidean space controllability of control systems with delay", J. SIAM Control Vol. 12, No. 4 (1974), 609-623.
3. R. B. Zmood and N. H. McClamrock, "On the pointwise completeness of differential-difference equations", J. Diff. Eqs. Vol. 12 (1972), 474-486.
4. A. Thowsen and W. R. Perkins, "On the controllability of linear time delay systems with piecewise constant inputs", to appear in International Jour. of Systems Science.
5. A. Olbrot, "Observability tests for constant time-lag systems", Control and Cybernetics Vol. 4 (1975), 71-84.
6. E. B. Lee, "Linear hereditary control systems", Calculus of Variations and Control Theory, Acad. Press Inc. (1976), 47-72.
7. A. Manitius, "Optimal control of hereditary systems", Autumn Math. course on Control Theory and Topics in Functional Analysis, Trieste, Sept. 1974.
8. J. Warga, "The reduction of certain control problems to an 'ordinary differential' type", SIAM Review Vol. 10 (1968), 219-222.
9. I. Ozhiganova, "A necessary and sufficient condition of one optimum problem for a linear differential equation with delay", Institut Khimicheskogo Mashinostroyeniya, Trudy., Vol. 25 (1963) 172-180.
10. A. W. Olbrot, "Control of retarded systems with function space constraints Part I: Necessary optimality conditions", to appear in Control and Cybernetics 1976.
11. A. Korytowski, "Optimal control for functional-differential systems with quadratic cost and constant, commensurable delays", Inst. Informatyki i Automatyki, AGH Krakow, 1976.

STATE AND INPUT FEEDBACK IN SYSTEMS CONTAINING TIME DELAYS*

Edward W. Kamen
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

Abstract

An operator theory is developed for the study of state and input feedback in systems described by a first-order delay differential equation. Necessary and sufficient conditions are given for coefficient assignability of the closed-loop systems's characteristic function. Coefficient assignability implies that the eigenvalues of the closed-loop system can be made finite in number and can be placed anywhere in the complex plane.

1. INTRODUCTION

In this paper we present an algebraic theory of feedback control for the class of systems given by a first-order delay differential equation of the form

$$\frac{dx(t)}{dt} = F_0 x(t) + \sum_{i=1}^{r} F_i x(t-a_i) + g_0 u(t) + \sum_{i=1}^{s} g_i u(t-b_i) \qquad (1)$$

In (1) the a_i and b_i are positive real numbers, the F_i (resp. g_i) are $n \times n$ (resp. $n \times 1$) matrices over the reals R, $x(t) \in R^n$ is the "instantaneous state", and $u(t) \in R$ is the input or control.

The feedback control of systems given by (1) has been studied extensively (see the survey [1]). In particular, there exists a deep theory on feedback controllers, optimal with respect to a quadratic cost functional, with optimal feedback "gains" obtainable from a Riccati-type equation (see Chapter IV of [2]).

Most of the previous work dealing with the control of (1) is based on expressing the solutions of (1) in terms of an evolution equation in a Hilbert or Banach space. This functional-analytical approach has yielded many results on the existence of feedback controllers that satisfy some performance criterion. However since the theory is based on Hilbert or Banach spaces, in general the expressions describing the feedback controllers can be solved for the feedback gains only with use of approximation methods (e.g. see [3]).

Although this approach may yield efficient designs, for some problems it is often possible to construct solutions in a much more simplified and direct manner by appealing to the algebraic structure of (1). We shall substantiate this claim in the present paper by considering the problem of eigenvalue assignment via state and input feedback.

In contrast to the functional-analytical approach, the theory presented below is based on the expression of (1) as a vector differential equation defined over a ring of delay operators. The construction of the operator ring framework (which is taken from [4,5]) is carried out in the next section.

---

*This work was supported by the U. S. Army Research Office, Research Triangle Park, NC, under Grant DAAG29-76-G-0048.

2. STATE FEEDBACK IN AN OPERATOR SETTING

Let δ_a with $a \in R$ denote the Dirac distribution (unit impulse) concentrated at the point $\{a\}$. Let D denote the set of all finite sums $\sum_{i=0}^{q} c_i \delta_{a_i}$ with $c_i \in R$ and $a_i \geq 0$. The set D is a commutative ring with the obvious addition and with convolution

$$\left(\sum_i c_i \delta_{a_i}\right) * \left(\sum_i d_i \delta_{b_i}\right) = \sum_j \sum_i c_j d_i \delta_{a_j + b_i}$$

Note that D contains $R\delta_0 = \{a\delta_0 : a \in R\}$, so that D can be viewed as a ring extension of the field R.

Let L_+^{loc} denote the space of R-valued Lebesgue-measurable locally-integrable functions (defined a.e. on R) with support bounded on the left. Support bounded on the left means that for any $v \in L_+^{loc}$, there exists a t_0, which in general depends on v, such that $v(t) = 0$ for all $t < t_0$. The set L_+^{loc} is also a ring with pointwise addition, $(u + v)(t) = u(t) + v(t)$, and with convolution

$$(u * v)(t) = \int_{-\infty}^{\infty} u(t - s)v(s)ds$$

Given $z = \sum_i c_i \delta_{a_i} \in D$ and $u \in L_+^{loc}$, we define the convolution

$$(z * u)(t) = \sum_i c_i u(t - a_i) \in L_+^{loc} \tag{2}$$

Now given (1) with $u \in L_+^{loc}$ and viewing x as an element in $(L_+^{loc})^n$ = space of n-element column vectors over L_+^{loc}, using (2) we can express (1) as a vector differential equation over the ring D

$$\frac{dx(t)}{dt} = (F * x)(t) + (g * u)(t) \tag{3}$$

where $F = F_0 \delta_0 + \sum_{i=1}^{r} F_i \delta_{a_i}$ and $g = g_0 \delta_0 + \sum_{i=1}^{s} g_i \delta_{b_i}$.

For the system (3), we shall first consider state feedback by setting $u = -k * x$ where k is a n-element row vector defined over a ring extension of D constructed as follows.

Let L_0 denote the subring of L_+^{loc} consisting of all functions with bounded support contained in $[0, \infty)$. Let J denote the set of all sums $z + v$ with $z \in D$ and $v \in L_0$. The set J is a ring with the obvious addition and with convolution

$$(z + v)*(y + w) = z*y + z*w + y*v + v*w, \quad z,y \in D, \quad v,w \in L_0$$

where $z*y$ is convolution in D, $z*w$ and $y*v$ are defined by (2), and $v*w$ is convolution in L_0. By construction, D and L_0 are subrings of J.

Given $e = z + v \in J$ and $u \in L_{+}^{loc}$, with $z = \sum_i c_i \delta_{a_i}$ and $v \in L_0$, we define the operation $e * u \in L_{+}^{loc}$ by

$$(e * u)(t) = \sum_i c_i u(t - a_i) + (v * u)(t) \tag{4}$$

Since the support of $v \in L_0$ is contained in $[0,a]$ for some $a > 0$, the convolution $v * u$ in (4) can be written in the form

$$(v * u)(t) = \int_{t-a}^{t} v(t - s)u(s)ds \tag{5}$$

Since the value of $v * u$ at t is the integral of $u(s)$ over the past interval $[t-a,t]$, the operation (5) is usually referred to as a distributed time delay.

Now let us consider state feedback by setting

$$u = - k * x + r \tag{6}$$

where $k = (k_1 \ldots k_n)$, $k_i \in J$, and r is an external input or disturbance. The operation $k * x$ in (6) is the usual multiplication of a row vector and a column vector with the componentwise operations defined by (4). Note that since the elements of k belong to J, we are allowing the feedback components (i.e. the k_i) to be pure and/or distributed time delays.

Combining (3) and (6), we get the following closed-loop equation

$$\frac{dx}{dt} = (F - g*k)*x + g*r \tag{7}$$

We shall study the closed-loop system (7) by employing the generalized derivative operator. This approach is based on the following constructions.

For $i = 0,1,2,\ldots$, let p^i denote the i^{th} generalized derivative of $\delta_0 (p^0 = \delta_0)$. For any $e \in J$, the convolution $e*p^i$ (defined in a space of distributions) is equal to the i^{th} generalized derivative of e (see [4]). Let $J[p]$ denote the set of all finite sums $\sum_{i=0}^{q} e_i * p^i$ with $e_i \in J$. The set $J[p]$ is contained in a space of distributions, but we do not need to consider this here. A key point is that $J[p]$ is a commutative ring with the following operations

$$\sum_i e_i * p^i + \sum_i f_i * p^i = \sum_i (e_i + f_i) * p^i$$

$$\left(\sum_i e_i * p^i\right) * \left(\sum_i f_i * p^i\right) = \sum_j \sum_i e_j * f_i * p^{j+i}$$

It is very important to note that an element of $J[p]$ does not have a unique expression as a "polynomial in p with coefficients in J." As we shall see later, this property plays a major role in the problem of eigenvalue placement via feedback.

Given the closed-loop system (7), consider the matrix $pI - F + g*k$ where I is the n×n identity matrix. Since $pI - F + g*k$ is a n×n matrix over the commutative ring $J[p]$, we can compute the determinant of $pI - F + g*k$, denoted by $\det(pI - F + g*k)$. The determinant of $pI - F + g*k$ is an element of $J[p]$ of the form

$$\det(pI - F + g*k) = p^n + \sum_{i=0}^{n-1} e_i * p^i, \quad e_i \in J \tag{8}$$

To make contact with the functional-analytical theory of (7), note that the Laplace transform of $\det(pI - F + g*k)$ is equal to the characteristic function of (7). Using (8), the Laplace transform of $\det(pI - F + g*k)$ is equal to $\sum_i \bar{e}_i(s)s^i$ where $\bar{e}_i(s)$ is the Laplace transform of e_i.

It is well known [6] that the asymptotic behavior of the solutions of (7) is determined by the location of the eigenvalues, which are the zeros of the characteristic function. In particular, (7) is asymptotically stable if and only if all eigenvalues have negative real parts. This result has been utilized by several authors (e.g. see [7] and [8]) to derive necessary and sufficient conditions for stabilizability by state feedback in cases for which g contains no delays.

We shall derive conditions for eigenvalue placement by using the ring structure on $J[p]$. We begin with the following concept.

Definition: Let F and g be over N where N is a subring of J with $R\delta_0 \subseteq N$. The element $\det(pI - F + g*k)$ is coefficient assignable with respect to N if for any $e_0, e_1, \ldots, e_{n-1}$ belonging to N, there exists an n-element row vector k over N such that

$$\det(pI - F + g*k) = p^n + \sum_{i=0}^{n-1} e_i * p^i$$

By choosing the e_i in the above definition to be of the form $c_i\delta_0$, $c_i \in R$, we see that coefficient assignability implies that the eigenvalues of the closed-loop system can be made finite in number and can be placed anywhere in the complex plane (with complex eigenvalues appearing in complex-conjugate pairs). Thus it would seem that conditions for coefficient assignability must be very strong, but such is not the case as revealed in the following section.

3. CONDITIONS FOR COEFFICIENT ASSIGNABILITY

We shall first give the following sufficient condition for coefficient assignability.

Theorem 1: Let F and g be over $N \subseteq J$ with $R\delta_0 \subseteq N$. Then $\det(pI - F + g*k)$ is coefficient assignable with respect to N if the determinant of $(g, F*g, \ldots, F^{n-1}*g)$ is an invertible element of N, where F^i is the i-fold convolution of F.

Proof: The proof follows by a straightforward generalization of the classical pole-placement theorem [9]. The constructions are as follows.

Write $\det(pI - F) = p^n + \sum_{i=0}^{n-1} z_i * p^i$, where $z_i \in D$, and define

$$\hat{F} = \begin{bmatrix} 0 & \delta_0 & 0 & \cdots & 0 \\ 0 & 0 & \delta_0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ & & & & \delta_0 \\ -z_0 & -z_1 & -z_2 & \cdots & -z_{n-1} \end{bmatrix} \qquad \hat{g} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \delta_0 \end{bmatrix}$$

Let $C = (g, F*g, \ldots, F^{n-1}*g)$ and suppose that det C is an invertible element of N. Then C has inverse C^{-1} over N. Let $\hat{C} = (\hat{g}, \hat{F}*\hat{g}, \ldots, \hat{F}^{n-1}*\hat{g})$ and define $T = \hat{C}*C^{-1}$. Then $\hat{g} = T*g$ and $\hat{F} = T*F*T^{-1}$. The pair $(\hat{F}, \hat{g})$ is called the control canonical form of (F,g). Now given $e_0, e_1, \ldots, e_{n-1}$ belonging to N, as a result of the special form of $(\hat{F}, \hat{g})$ we have that

$$\det(pI - \hat{F} + \hat{g}*\hat{k}) = p^n + \sum_{i=0}^{n-1} e_i * p^i$$

where $\hat{k} = (e_0 - z_0 \quad e_1 - z_1 \quad \cdots \quad e_{n-1} - z_{n-1})$.
Defining $k = \hat{k}*T$, we have that

$$\det(pI - F + g*k) = p^n + \sum_{i=0}^{n-1} e_i * p^i \quad \blacksquare$$

Unfortunately, the condition in Theorem 1 is very strong except when $N = R\delta_0$, the finite-dimensional case. For example, if N = D, then $\det(g, F*g, \ldots, F^{n-1}*g)$ is an invertible element of D if and only if it is equal to $a\delta_0$ where a is a nonzero element of R.

We may then ask if there is a "weaker" sufficient condition. The answer is no if p is transcendental over N which means that the elements of N[p] have unique expressions as polynomials in p with coefficients in N. If p is transcendental over N, the condition in Theorem 1 is a necessary and sufficient condition for coefficient assignability with respect to N. The proof of necessity is given in [5,Theorem 11].

Given positive real numbers $a_1, a_2, \ldots, a_r$, let $R[\delta_{a_1}, \ldots, \delta_{a_r}]$ denote the smallest subring of D containing δ_{na_i}, n = 0,1,2,... and i = 1,2,...,r. If the delays $a_1, \ldots, a_r$ are noncommensurate, then p is transcendental over $N = R[\delta_{a_1}, \ldots, \delta_{a_r}]$ (see [4,Theorem 1]). Hence the condition in Theorem 1 above is a necessary and sufficient condition for coefficient assignability with respect to $R[\delta_{a_1}, \ldots, \delta_{a_r}]$. Note that the elements of $R[\delta_{a_1}, \ldots, \delta_{a_r}]$ can be realized by parallel interconnections of scalors and pure time delays.

This result implies that for systems with pure delays, it is seldom possible to achieve coefficient assignability using feedback components belonging to $R[\delta_{a_1},\ldots,\delta_{a_r}]$. However this is not the case if we allow the feedback components to contain distributed time delays (i.e. let $N = J$).

If $N = J$, then p is not transcendental over N and the condition in Theorem 1 is not necessary. For $N = J$, we have the following necessary and sufficient condition for coefficient assignability.

Theorem 2: Given F and g defined over D, $\det(pI - F + g*k)$ is coefficient assignable with respect to J if and only if there exists a $n\times n$ matrix W over J such that

$$W*[\mathrm{cof}(pI - F)]^{TR}*g = (\delta_0 \quad p \quad p^2 \ \ldots \ p^{n-1})^{TR} \tag{10}$$

where $[\mathrm{cof}(pI - F)]^{TR}$ is the transpose of the matrix of cofactors of $pI - F$.

Proof: The proof, first given in [5], is based on the determinant identity

$$\det(pI - F + g*k) = \det(pI - F) + k*[\mathrm{cof}(pI - F)]^{TR}*g \tag{11}$$

Again, write $\det(pI - F) = p^n + \sum_{i=0}^{n-1} z_i*p^i$, $z_i \in D$, and suppose that there exists a W over J satisfying (10). Given $e_0, e_1, \ldots, e_{n-1}$ belonging to J, let

$$k = (e_0 - z_0 \quad e_1 - z_1 \quad \ldots \quad e_{n-1} - z_{n-1})*W \tag{12}$$

Inserting (12) into (11) and using (10), we have that

$$\det(pI - F + g*k) = p^n + \sum_{i=0}^{n-1} e_1*p^i$$

Thus (10) is a sufficient condition for coefficient assignability with respect to J. The necessity of condition (10) also follows from (11). ■

If $\det(g, F*g, \ldots, F^{n-1}*g)$ is an invertible element of J so that $C = (g, F*g, \ldots, F^{n-1}*g)$ has an inverse over J, then (10) is satisfied with $W = \hat{C}*C^{-1}$ where $\hat{C}$ is the $n\times n$ matrix defined in the proof of Theorem 1. However, when $\det C$ is not an invertible element of J, it is still often possible to satisfy (10). We illustrate this for the following class of systems.

Suppose that we are given the following delay differential equation

$$\frac{dx(t)}{dt} = Ax(t - a) + bu(t) \tag{13}$$

where A (resp. b) is a $n\times n$ matrix (resp. n-element column vector) over the reals R and $a > 0$. Note that in terms of the operator notation (see (3)), we have that $F = A\delta_a$ and $g = b\delta_0$.

Theorem 3: $\det(pI - A\delta_a + b\delta_0*k)$ is coefficient assignable with respect to J if $\det(b, Ab, \ldots, A^{n-1}b)$ is nonzero.

A constructive proof of Theorem 3 is given in [10]. This result shows that the system (13) is "almost always" coefficient assignable with respect to J. What is happening here is that by using feedback components containing distributed time delays, we can insure that the pure time delay $x(t - a)$ does not appear in the closed-loop equation.

4. STATE AND INPUT FEEDBACK

When g in the system equation (3) is equal to $\bar{g}\delta_a$, $\bar{g} \in D^n$, the condition (10) can not be satisfied. However in such cases we may still be able to achieve coefficient assignability by considering state and input feedback.

We can feed back the state and input by setting

$$u = -v * u - k * x + r \tag{14}$$

where $v \in L_0$ and k is a n-element row vector over J. Combining (3) and (14), we have the following closed-loop equation

$$(\delta_0 + v) * \frac{dx}{dt} = (\delta_0 + v) * F * x - g * k * x + g * r \tag{15}$$

Note that (15) is a functional differential equation of the neutral type.

Let Q denote the quotient field of J; that is, Q consists of all ratios e/f with $e, f \in J$, $f \neq 0$, with the operations $e_1/f_1 + e_2/f_2 = (e_1+e_2)/(f_1+f_2)$ and $(e_1/f_1)(e_2/f_2) = (e_1 * e_2)/(f_1 * f_2)$. We shall sometimes denote δ_0/f by f^{-1}. Given $e/f \in Q$, define $(e/f)p = (e*p)/(f*p)$, so that we can construct the ring Q[p] consisting of all finite sums $\sum_{i=0}^{q} h_i p^i$, $h_i \in Q$.

Then given (15), consider the n×n matrix $pI - F + (\delta_0 + v)^{-1} g*k$, which is defined over Q[p]. The determinant of this matrix is an element of Q[p] of the form

$$\det(pI - F + (\delta_0 + v)^{-1} g*k) = p^n + \sum_{i=0}^{n-1} h_i p^i, \ h_i \in Q.$$

Definition: $\det(pI - F + (\delta_0+v)^{-1} g*k)$ is coefficient assignable (up to a multiplicative factor) if for any $e_0, e_1, \ldots, e_{n-1}$ belonging to J, there exist $v \in L_0$ and $k = (k_1 \ \ldots \ k_n)$, $k_i \in J$, such that

$$(\delta_0+v) * \det(pI - F + (\delta_0+v)^{-1} g*k) = p^n + \sum_{i=0}^{n-1} e_i * p^i \tag{16}$$

We shall reveal the significance of this concept after giving the following sufficient condition for coefficient assignability.

Theorem 4: $\det(pI - F + (\delta_0+v)^{-1} g*k)$ is coefficient assignable if there exists a n×(n+1) matrix $W = (w_1 \ \ w_2 \ \ldots \ w_n \ \ w_{n+1})$

with $w_i \in J^n$ for $i = 1,2,\ldots,n$ and $w_{n+1} \in L_0^n$, such that

$$W * \begin{bmatrix} [\mathrm{cof}(pI - F)]^{TR} * g \\ \\ \\ \det(pI - F) \end{bmatrix} = \begin{bmatrix} \delta_0 \\ p \\ p^2 \\ \vdots \\ p^{n-1} \end{bmatrix} \tag{17}$$

Proof: Write $\det(pI - F) = p^n + \sum_i z_i * p^i$, and suppose that there exists a $W = (w_1 \quad w_2 \ldots w_n \quad w_{n+1})$ satisfying (17). Given $e_0, e_1, \ldots, e_{n-1}$ belonging to J, let

$$(k \quad v) = (e_0 - z_0 \quad e_1 - z_1 \ldots e_{n-1} - z_{n-1}) * W \tag{18}$$

Using (17),(18) and a modified version of (11), we get (16). ■

Again, write $\det(pI - F) = p^n + \sum_i z_i * p^i$, and suppose that $W = (w_1 \ \ldots. \ w_n \quad w_{n+1})$ satisfies (17). Suppose also that there exist real numbers $b_1, b_2, \ldots, b_n$ such that

(i) the zeros of $p^n + \sum_i b_{i+1} p^i$ are in the left half of the complex plane;

(ii) the following condition holds

$$\int_0^\infty \left| \sum_{i=1}^n (b_i w_{i,n+1}(t) - (z_i * w_{i,n+1})(t)) \right| dt < 1$$

where $w_{i,n+1}$ is the i^{th} component of w_{n+1}.

Condition (ii) is needed to insure that the component defined by $(\delta_0 + v)^{-1}$ is input/output stable. It follows from Theorem 7 in [5] that with k and v given by (18) and with $e_i = b_{i+1}\delta_0$, the closed-loop system (15) is asymptotically stable and the limit of u(t), given by (14), as t approaches infinity is equal to zero for all initial functions. Thus coefficient assignability, along with conditions (i,ii), implies that the system (3) can be stabilized using state and input feedback. The complete version of this theory will be presented in [10].

Example: Consider the following system

$$\frac{dx_1(t)}{dt} = x_2(t)$$

$$\frac{dx_2(t)}{dt} = u(t - a), \ a > 0$$

For this example,

$$F = \begin{bmatrix} 0 & \delta_0 \\ 0 & 0 \end{bmatrix} \quad \text{and } g = \begin{bmatrix} 0 \\ \delta_a \end{bmatrix}$$

$$\left[\mathrm{cof}(pI - F)\right]^{TR} * g = \begin{bmatrix} \delta_a \\ \delta_a p \end{bmatrix}$$

and

$$\det(pI - F) = p^2$$

Solving

$$W * \begin{bmatrix} \delta_a \\ \delta_a p \\ p^2 \end{bmatrix} = \begin{bmatrix} \delta_0 \\ p \end{bmatrix}$$

we get the following nonunique solution

$$W = \begin{bmatrix} \delta_0 & h & h*h \\ 0 & \delta_0 & h \end{bmatrix}$$

where $h(t) = 1$ for $0 < t < a$ and $h(t) = 0$ for all other t. Thus the closed-loop system is coefficient assignable.

Given real numbers b_1 and b_2, using (18) with $e_i = b_{i+1}\delta_0$, we have that

$$k = (b_1\delta_0 \qquad b_1 h + b_2\delta_0)$$

and

$$v = h*(b_1 h + b_2\delta_0) \tag{19}$$

Now condition (i) is satisfied if b_1 and b_2 are positive real numbers and (ii) is satisfied if

$$\int_0^\infty \left|v(t)\right| dt < 1$$

Using (19), we have

$$\int_0^\infty \left|v(t)\right| dt \le \int_0^a b_1 h(t)\,dt \quad + ab_2$$

$$\le ab_1 + ab_2 = a(b_1 + b_2)$$

Thus we can pick any positive numbers b_1 and b_2 such that

$$b_1 + b_2 \le 1/a$$

REFERENCES

1. H.T. Banks and A. Manitius, "Application of abstract variational theory to hereditary systems - a survey," IEEE Trans. on Automatic Control,Vol. AC-19,1974,pp. 524-533.

2. A. Manitius, "Optimal control of hereditary systems," Research Report CRM-472,Centre de Recherches Mathématiques, Université de Montréal,1974.

3. M.C. Delfour, "Numerical solution of opt. contr. prob. for linear hereditary systems with a linear-quadratic cost function and approximation of the Riccati differential equation," Research Report CRM-408,Centre de Recherches Mathématiques, Université de Montréal,1974.

4. E.W. Kamen,"On an algebraic theory of systems defined by convolution operators,"Math. Systems Theory,Vol. 9,No. 1, 1975, pp.57-74.

5. E.W. Kamen,"An operator theory of linear functional differential equations,"submitted to J. Diff. Equ.

6. J.K. Hale,"Functional differential equations," Springer-Verlag, New York,1971.

7. H.F. Vandevenne,"Controllability and stabilizability properties of delay systems,"in Proc. 1972 IEEE Conf. Decision and Control,pp.370-377.

8. L.Pandolfi,"Feedback stabilization of functional differential equations,' to appear in Boll. Un. Mat. It.

9. R.E. Kalman,"Lectures on controllability and observability," CIME Summer Course, Cremonese, Roma, 1969.

10. E.W. Kamen,"Feedback control of time-delay systems," in preparation.

FUNCTION SPACE CONTROLLABILITY OF RETARDED SYSTEMS: SOME NEW ALGEBRAIC CONDITIONS*

A. MANITIUS
Centre de Recherches Mathématiques
Université de Montréal
Montréal, Canada, H3C 3J7

ABSTRACT

The paper gives an extension of earlier results of [8] on M_2-approximate controllability of retarded systems. The role of spectral controllability is analyzed. A necessary and sufficient condition for completeness of generalized eigenfunctions associated with the retarded equation is given. Two sufficient conditions for M_2-approximate controllability based on spectral controllability assumption are given along with some examples.

1. INTRODUCTION

The function space controllability of linear retarded systems has been recently investigated from the point of view of the theory of abstract linear systems [8]. This led to new necessary and sufficient conditions for M_2- and L_2-approximate controllability expressed in terms of system's transfer function. In case of systems with single delay these conditions were reduced to some purely algebraic tests involving large matrices.

Continuing these investigations, in this note we consider the relationship between the spectral controllability and the M_2-approximate controllability. It was shown in [8] that the spectral controllability is a necessary condition for M_2-approximate controllability. For systems with single delay the test for spectral controllability is quite simple. If, therefore, one could prove that, perhaps under some additional hypotheses, the spectral controllability is also sufficient for M_2-approximate controllability, one would have a simple test for the latter.

Notation in this paper follows closely that of [8,9,10]. The class of systems considered is

$$\dot{z}(t) = A_0 z(t) + A_1 z(t-h) + Bu(t) \tag{1.1}$$

with $z \in R^n$, $u \in R^m$. Define $\Delta(\lambda,\mu) = I\lambda - A_0 - A_1\mu$, $\Delta(\lambda) \equiv \Delta(\lambda,e^{-\lambda h})$. As before, $P(\lambda)$ denotes the n×m polynomial matrix defined by the identity

$$\text{adj } \Delta(\lambda,\mu)B = P(\lambda)v(\mu) \tag{1.2}$$

where $v(\mu) = (1,\mu,\mu^2,\dots,\mu^{n-1})^T \otimes I_m$, "adj" is the matrix adjoint, I_m is an m×m identity matrix, ⊗ denotes the Kronecker product. $S(t)$ denotes the usual semigroup operator in M_2 [1,3], A its infinitesimal generator and $\sigma(A)$ - the spectrum of A, i.e. the set of all the roots of det $\Delta(\lambda)$. The M_2-, L_2-approximate controllability and spectral controllability are defined in [8,9,10,4].

Because of the lack of space, the results will be announced without proofs, which will appear elsewhere. Other contributions to controllability are surveyed in [8] and [6]. We collect below for an easy reference some of previously obtained conditions [8].

(C1) A necessary and sufficient condition for M_2-approximate controllability of (2.1) is that the identity

---

* This research was supported in part by the National Research Council of Canada under Grant A-9240 and in part by the Ministère de l'Education du Gouvernement du Québec under Grant FCAC 1975/76.

$$(1.3) \qquad [c+q(\lambda)]^T P(\lambda) \equiv w_M(\lambda)$$

implies $c=0$, $q(\lambda) \equiv 0$ (hence $w_M(\lambda) \equiv 0$), where $c \in R^n$;

(1.4) $q(\lambda)$ is a finite Laplace transform of an n-vector valued L_2-function defined on $[0,h]$;

$$(1.5) \qquad w_M(\lambda) = [\gamma(\lambda),0]\, e^{-\lambda h} - [0,\gamma(\lambda)];$$

$$(1.6) \qquad \gamma(\lambda) = [p_{n-2}(\lambda),\ldots,p_1(\lambda),p_0]$$

and p_k are some polynomials of degree at most k.

(C2) A necessary condition for M_2-approximate controllability is that the rank of the polynomial matrix $P(\lambda)$ be n.

(C3) A necessary and sufficient condition for spectral controllability in case m=1 is $P(\lambda)v(e^{-\lambda h}) \neq 0 \;\; \forall\, \lambda$.

2. COMPLETENESS OF EIGENFUNCTIONS AND CONTROLLABILITY

For $\lambda \in \sigma(A)$ let $\mathcal{M}_\lambda$ denote a generalized eigenspace of A, and let span $\mathcal{M}_\lambda$ denote the linear manifold spanned by all $\mathcal{M}_\lambda$, $\forall\, \lambda \in \sigma(A)$. The system of generalized eigenfunctions of A is said to be complete in the space M_2 if

$$(2.1) \qquad \overline{\text{span}}\; \mathcal{M}_\lambda = M_2$$

where the bar denotes closure in M_2-topology.

If the above relation holds, then the spectral controllability is sufficient for M_2-approximate controllability, because it implies

$$(2.2) \qquad \text{span}\; \mathcal{M}_\lambda \subset K_\infty \overset{df}{=} \bigcup_{t>0} K_t$$

where K_t is the attainable set in M_2 from the origin, by using, say, L_2-controls. Consequently, by (2.1) and (2.2) one has

$$(2.3) \qquad \overline{K}_\infty \supset \overline{\text{span}}\; \mathcal{M}_\lambda = M_2$$

so that $\overline{K}_\infty = M_2$, as required.

Some sufficient conditions for completeness in case of scalar equations were given in [7]. We present below a more general result, valid for any n-vector retarded system of a class considered in [1,4], (not just for simple system 1.1).

Theorem 2.1. The system of generalized eigenfunctions of A is complete in M_2 if and only if there is no nonzero pair $(c,q(\lambda))$, $c \in R^n$, $q(\lambda)$ characterized by (1.4), such that the function $[c+q(\lambda)]^T \Delta^{-1}(\lambda)$ is entire.

For systems with single delay (1.1), by making use of the Pragmén-Lindelöf theorem one can prove that the condition given by Theorem 2.1 is satisfied if

$$(2.4) \qquad \text{rank } A_1 = n.$$

Corollary 2.2. If (2.4) holds, then the condition $P(\lambda)v(e^{-\lambda h}) \neq 0 \;\; \forall\, \lambda$ is sufficient for M_2-approximate controllability.

In the event the completeness does not hold, the space $\overline{\text{span}\ M_\lambda}$ is a proper subspace of M_2, but the evolution of the retarded system on some initial interval $[0,\delta]$, $\delta \le nh$, is not necessarily restricted to $\overline{\text{span}\ M_\lambda}$ (see e.g. [5, Corollary 2] and [2, p. 316]). Generally speaking, therefore, the lack of completeness of span M_λ in M_2 does not necessarily prevent the M_2-approximate controllability.

Example 2.1. Consider the system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t-h) \\ x_2(t-h) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t).$$

The spectrum consists of a double eigenvalue $\lambda = 0$; the span M_λ is at most 2-dimensional and, therefore, incomplete in M_2. However, since $A_0 = 0$ and (A_1,B) is controllable, the system is M_2-approximately controllable (see [8,9]).

As a side observation we add that the pair $(c,q(\lambda))$ referred to in Theorem 2.1 is, in this case

$$c^T = [1,0] \qquad q^T(\lambda) = [(e^{-\lambda}-1)/\lambda,\ 0]$$

so that

$$[c^T+q^T(\lambda)]\Delta^{-1}(\lambda) = [(\lambda+e^{-\lambda}-1)/\lambda^2,\ 0].$$

This function has a complete cancellation of poles by zeros, hence is entire.

This indicates that the completeness, while convenient, is too strong a hypothesis. To find some weaker conditions, we will assume that (C2) and (C3) hold but (C1) fails. To make use of these assumptions, we begin with some additional results on the properties of $P(\lambda)$.

3. RIGHT ANNIHILATORS OF P(λ) AND THE SUFFICIENT CONTROLLABILITY CONDITIONS

Consider the points λ at which the rank of the numerical matrix $P(\lambda)$ is less than n (in case $m=1$ such points λ satisfy $\det P(\lambda) = 0$).

Theorem 3.1. Let the polynomial matrix $P(\lambda)$ have rank n. For each λ_0 such that $\text{rank}\ P(\lambda_0) < n$ and $\det(I\lambda_0-A_0) \neq 0$ there exists a $\mu \neq 0$ such that $\det \Delta(\lambda_0,\mu) = 0$ and $P(\lambda_0)v(\mu) = 0$. If $\det(I\lambda_0-A_0) = 0$ then the conclusion is still valid for $m=1$, except that now μ can be zero: $\mu=0$ if and only if $\text{rank}[I\lambda_0-A_0,B] < n$.

Example 3.1. Let $n=2$, $m=1$, and

$$A_0 = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix} \qquad A_1 = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \qquad B = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

We find $\det \Delta(\lambda,\mu) = (\lambda-1+\mu)(\lambda-2-\mu)$

$$P(\lambda) = \begin{bmatrix} \lambda & -1 \\ 2(\lambda-1) & 2 \end{bmatrix}$$

$$\det P(\lambda) = 4\lambda-2$$

$$\lambda_0 = \tfrac{1}{2}.$$

It turns out that $\det \Delta(\lambda_0,\mu) = (\mu - \frac{1}{2})(-\mu - \frac{3}{2}) = 0$ gives two possible values for μ. Substituting $\lambda_0 = \frac{1}{2}$ into $P(\lambda)$ we see that

$$\begin{bmatrix} \frac{1}{2} & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ \mu \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{for } \mu = \frac{1}{2}$$

(the fact that $\mu = \lambda_0$ is just a coincidence).

In case m=1 one can also prove that if λ_0 is a multiple root of det $P(\lambda)$, then there exists a polynomial $\mu(\lambda)$ and an integer $k \geq 0$ such that, defining the polynomial vector $z(\lambda) = v(\mu(\lambda))$, one has

$$(3.1) \qquad \frac{1}{i!} \frac{d^i}{d\lambda^i}\bigg|_{\lambda_0} [P(\lambda)z(\lambda)] = 0 \qquad i=0,1,\ldots,k.$$

Because of the lack of space, in this note we limit the description to the case of m=1 with single roots of det $P(\lambda)$, although it was verified that, by using (3.1), the whole subsequent analysis extends also to the multiple root case.

From (1.3) we obtain

$$(3.2) \qquad c^T + q^T(\lambda) = w_M(\lambda)P^{-1}(\lambda).$$

In order that $q(\lambda)$, hence also $c+q(\lambda)$, be an entire function, all the poles of $P^{-1}(\lambda)$ must be cancelled by premultiplication of $P^{-1}(\lambda)$ by $w_M(\lambda)$.

It is instructive to consider first the case n=2. Then det $P(\lambda)$ is of degree at most one. If it is of degree zero, then $P^{-1}(\lambda)$ is a polynomial matrix, and from (3.2) one has $q^T(\lambda) = w_M(\lambda)P^{-1}(\lambda) - c^T$ = polynomial in λ and $e^{-\lambda h}$; this, however, combined with (1.4) implies [8], by the Paley-Wiener theorem, that $q(\lambda) \equiv 0$; consequently $w_M(\lambda) \equiv 0$ and so c=0. If det $P(\lambda)$ is of degree one, we take its root λ_1 and a corresponding number μ_1 referred to in Theorem 3.1, and multiply both sides of (1.3) by the vector $[1,\mu_1]^T$. This yields

$$0 = p_0[e^{-\lambda_1 h}, -1][1,\mu_1]^T,$$

hence either $p_0 = 0$ or $\mu_1 = e^{-\lambda_1 h}$. But the latter is excluded by (C3). Hence p_0 must be zero, $w_M(\lambda) \equiv 0$, and, by (3.2), $c+q(\lambda) \equiv 0$ which is possible only [8] when c=0 and $q(\lambda) \equiv 0$. We have thus proved

Corollary 3.2. For $n \leq 2$, $m = 1$, spectral controllability is both necessary and sufficient for M_2-approximate controllability.

For $n \geq 2$ the situation is more complicated, and it is not at present known whether the conclusion of Corollary 3.2 holds without some additional assumptions. We proceed further by taking one such assumption.

Let m=1 and let the pair (A_1,B) be controllable. Then det $P(\lambda)$ is a polynomial of degree $N = n(n-1)/2$. For each root of this polynomial λ_i define μ_i to be a corresponding number satisfying $P(\lambda_i)v(\mu_i) = 0$. The numbers μ_i can be found as the roots of det $M(\mu)$, where $M(\mu)$ is an $n \times n$ polynomial matrix (with det $M(\mu)$ of degree N, too) defined by the identity

$$(3.3) \qquad \text{adj } \Delta(\lambda,\mu)B = M(\mu)v(\lambda).$$

Assume that λ_i are simple and multiply Eq. (1.3) by $v(\mu_i)$, $i=1,\ldots,N$. This gives

$$\begin{aligned} 0 &= ([\gamma(\lambda_i),0]\,e^{-\lambda_i h} - [0,\gamma(\lambda_i)])v(\mu_i) \\ &= [p_{n-2}(\lambda_i),\ldots,p_0,0]v(\mu_i)(e^{-\lambda_i h}-\mu_i) \\ &= [p_{n-2}(\lambda_i)+p_{n-3}(\lambda_i)\mu_i+\ldots+p_0\mu_i^{n-2}](e^{-\lambda_i h}-\mu_i). \end{aligned} \tag{3.4}$$

Since $P(\lambda)v(e^{-\lambda h}) \neq 0$ excludes $\mu_i = e^{-\lambda_i h}$, the point (λ_i,μ_i) must be a root of the polynomial in two variables

$$g(\lambda,\mu) = p_{n-2}(\lambda) + p_{n-3}(\lambda)\mu + \ldots + p_0\mu^{n-2}. \tag{3.5}$$

Since each $p_j(\lambda)$, $j=0,\ldots,n-2$, can be written as

$$p_j(\lambda) = \sum_{\ell=0}^{j} \beta_{j\ell}\lambda^{\ell} \quad \text{where } \beta_{j\ell} \text{ are some coefficients,} \tag{3.6}$$

the number of undetermined coefficients in $g(\lambda,\mu)$ is

$$\sum_{j=0}^{n-2} (j+1) = n(n-1)/2 = N.$$

Assuming that the λ_i, μ_i have been found, the condition

$$g(\lambda_i,\mu_i) = 0 \qquad i=1,\ldots,N$$

reduces to a systems of N linear homogeneous equations in N unknowns $\beta_{j,\ell}$, $j=0,\ldots,n-2$, $\ell=0,\ldots,j$

$$\sum_{j=0}^{n-2}\left(\sum_{\ell=0}^{j} \beta_{j,\ell}\lambda_i^{\ell}\right)\mu_i^{n-2-j} = 0 \qquad i=1,\ldots,N. \tag{3.7}$$

The determinant of this system, after some rearrangements, is

$$D\begin{pmatrix}\lambda_1 & \cdots & \lambda_N \\ \mu_1 & \cdots & \mu_N\end{pmatrix} = \begin{vmatrix} 1 & \lambda_1 & \mu_1 & \lambda_1^2 & \lambda_1\mu_1 & \mu_1^2 & \lambda_1^3 & \cdots & \mu_1^{n-2} \\ 1 & \lambda_2 & \mu_2 & \lambda_2^2 & \lambda_2\mu_2 & \mu_2^2 & \lambda_2^3 & \cdots & \mu_2^{n-2} \\ \cdot & & & & & & & & \cdot \\ \cdot & & & & & & & & \cdot \\ \cdot & & & & & & & & \cdot \\ 1 & \lambda_N & \mu_N & \lambda_N^2 & \lambda_N\mu_N & \mu_N^2 & \lambda_N^3 & \cdots & \mu_N^{n-2} \end{vmatrix} \tag{3.8}$$

which is an analogy of the Vandermonde determinant. If this determinant is not zero, Eq. (3.7) implies $\beta_{j,\ell} = 0$, $j=0,\ldots,n-2$, $\ell=0,\ldots,j$, so that $p_j \equiv 0$, $j=0,\ldots,n-2$, and consequently $c^T = 0$, $q^T(\lambda) \equiv 0$, hence the M_2 approximate controllability holds.

Similar considerations can be repeated for multiple roots λ_i, with a suitable modification of $D(\cdot)$. Therefore, the following result is true:

Theorem 3.3. Let $m=1$, and assume

(i) the pair (A_1,B) is controllable

(ii) $P(\lambda)v(e^{-\lambda h}) \neq 0 \quad \forall\ \lambda$

(iii) $D\begin{pmatrix}\lambda_1 & \cdots & \lambda_N \\ \mu_1 & \cdots & \mu_N\end{pmatrix} \neq 0.$

Then the system is M_2-approximately controllable.

If (A_1,B) is not controllable, then $\deg \det P(\lambda) < N$, so that the number of equations (3.7) is smaller than the number of unknowns, so that the above reasoning does not apply directly.

Example 3.2. In this example the spectrum $\sigma(A)$ is finite but the system is M_2-approximately controllable.

$$A_0 = \begin{bmatrix} -3 & -2 & 0 \\ 0 & 0 & 2 \\ 0 & -1 & -3 \end{bmatrix} \qquad A_1 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \qquad B = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}.$$

First of all we notice that (A_1,B) is a controllable pair. We then compute successively $\Delta(\lambda)$, $P(\lambda)$, $M(\mu)$ and the roots of determinants of all three matrices

$$P(\lambda) = \begin{bmatrix} \lambda^2-\lambda-10 & -2 & 0 \\ 2(\lambda+3)^2 & \lambda+3 & 0 \\ -2(\lambda+3) & 2\lambda+5 & 1 \end{bmatrix} \qquad M(\mu) = \begin{bmatrix} -2\mu-10 & -1 & 1 \\ 3(\mu+6) & \mu+12 & 2 \\ (\mu-1)(\mu+6) & 2(\mu-1) & 0 \end{bmatrix}$$

$\det \Delta(\lambda) = (\lambda+3)(\lambda^2+3\lambda+2) \equiv \det P(\lambda)$. The roots $\lambda_1 = -3$, $\lambda_2 = -2$, $\lambda_3 = -1$ coincide with the eigenvalues of A, and the spectrum of A is finite. $\det M(\mu) = -(\mu-1)(\mu^2+6\mu+8)$. The condition $P(\lambda_i)v(\mu_i) = 0$ implies the following ordering of roots $\mu_i : \mu_1 = 1$, $\mu_2 = -2$, $\mu_3 = -4$. By verifying that

$$P(\lambda_i)v(e^{-\lambda_i h}) \neq 0 \qquad i=1,2,3,$$

one has that $P(\lambda)v(e^{-\lambda h}) \neq 0$ for $\forall\ \lambda$. We now compute from (3.8)

$$D\begin{pmatrix}\lambda_1 & \lambda_2 & \lambda_3 \\ \mu_1 & \mu_2 & \mu_3\end{pmatrix} = \begin{vmatrix} 1 & -3 & 1 \\ 1 & -2 & -2 \\ 1 & -1 & -4 \end{vmatrix} = 1.$$

Since all the conditions of Theorem 3.3 are fulfilled, the system is M_2-approximately controllable. Note that even though the pair (A_1,B) is controllable, the criterion for right triangularity of $P(\lambda)$ of [8,9] is not satisfied.

Remark 3.1. For three dimensional systems with single roots of $\det P(\lambda)$ and satisfying (i) (ii) one can show that the condition (iii) is redundant. Take (A_1,B) in the companion canonical form, and A_0 arbitrary, $A_0 = \{a_{ij},\ i,j = 1,2,3\}$. Then

$$\operatorname{adj} \Delta(\lambda,\mu)B = \begin{bmatrix} (\mu+a_{12})(\mu+a_{23}) + a_{13}(\lambda-a_{22}) \\ (\mu+a_{23})(\lambda-a_{11}) + a_{21}a_{13} \\ (\lambda-a_{11})(\lambda-a_{22}) - a_{21}(a_{12}+\mu) \end{bmatrix} = 0.$$

Since the roots $\lambda_1,\lambda_2,\lambda_3$ of det $P(\lambda)$ are distinct, $a_{21} \neq 0$. The third row of $P(\lambda_i)v(\mu_i) = \operatorname{adj} \Delta(\lambda_i,\mu_i)B = 0$ gives

$$(\lambda_i-a_{11})(\lambda_i-a_{22}) - a_{21}a_{12} - a_{21}\mu_i = 0 \qquad i=1,2,3.$$

Eq. (3.5) gives

$$(3.9) \qquad \beta_{1,0} + \beta_{1,1}\lambda_i + \beta_{0,0}\mu_i = 0 \qquad i=1,2,3.$$

Since λ_i are distinct, $\beta_{0,0}$ must be nonzero. Substituting this equation into the previous one we obtain

$$(\lambda_i-a_{11})(\lambda_i-a_{22}) - a_{21}a_{12} + \frac{a_{21}}{\beta_{0,0}}(\beta_{1,0}+\beta_{1,1}\lambda_i) = 0 \qquad i=1,2,3$$

which is impossible since there is no nontrivial second degree curve having three distinct zeros. Thus the equation (3.9) can be satisfied only if $\beta_{1,0} = \beta_{1,1} = \beta_{0,0} = 0$, which implies the M_2-approximate controllability.

Generally speaking, for systems with (A_1,B) controllable the equations $\operatorname{adj} \Delta(\lambda_i,\mu_i)B = 0$, $i=1,\dots,N$, and $g(\lambda_i,\mu_i) = 0$, $i=1,\dots,N$, state that n-algebraic curves of degree (n-1) in (λ,μ) and 1 algebraic curve of degree (n-2) have $N = n(n-1)/2$ common intersection points (λ_i,μ_i); however, the simple argument given above to exclude this possibility in case n=3 does not seem to generalize directly to cases with n>3.

We now turn briefly to the case of systems satisfying (C2) which are not L_2-approximately controllable (see [8; sections 3 and 8]). For $n\leq 2$ this situation cannot happen. For n=3 we have this result.

Theorem 3.5. Let m=1, n=3. If the system (1.1) satisfies (C2) but is not L_2-approximately controllable, then

(i) det $P(\lambda)$ is of degree 3, and its three roots, including possibly multiple roots, are roots of det $\Delta(\lambda)$. (det $P(\lambda)$ divides det $\Delta(\lambda)$.)

(ii) The transfer function $\Delta^{-1}(\lambda)B$ has three cancellations corresponding to the roots of det $P(\lambda)$; the system is not spectrally controllable. In particular if det $\Delta(\lambda)$ is a polynomial, then the transfer function $\Delta^{-1}(\lambda)b$ is entire (the system is a transpose of a pointwise degenerate system).

Corollary 3.6. For m=1, n=3 the conditions det $P(\lambda) \neq 0$ and $P(\lambda)v(e^{-\lambda h}) \neq 0$ guarantee the L_2-approximate controllability.

Proofs of these results are also based on Theorem 3.1.

These results and the results of [8, sec. 8] indicate that the lack of L_2-approximate controllability in systems whose matrix $P(\lambda)$ has rank n $(n\geq 3)$ (a necessary condition of L_2-approximate controllability, [8, sec. 3] is probably due to the existence of a 3×3, or larger, subsystem which is a transpose of a pointwise degenerate system.

REFERENCES

[1] H.T. Banks and J.A. Burns, Hereditary control problems: numerical methods based on averaging approximations, Research report, Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, 1976.

[2] H.T. Banks and A. Manitius, Projection series for functional differential equations with applications to optimal control problems, J. Differential Equations, 18 (1975), 296-332.

[3] C. Bernier and A. Manitius, L'étude des semi-groupes d'opérateurs associés aux équations linéaires retardées, Report CRM-623, Centre de Recherches Mathématiques, Université de Montréal, 1976.

[4] M.C. Delfour and S.K. Mitter, Controllability, observability and optimal feedback control of affine hereditary differential systems, SIAM J. Control, 10 (1972), No. 2, 298-328.

[5] D. Henry, Small solutions of linear autonomous functional differential equations, J. Differential Equations 8, (1970), 494-501.

[6] M.Q. Jacobs and C.E. Langenhop, Criteria for function space controllability of linear neutral systems, SIAM J. Control and Optimization (to appear).

[7] N. Levinson and C. McCalla, Completeness and independence of the exponential solutions of some functional differential equations, Studies in Applied Math. 53 (1974), 1-15.

[8] A. Manitius and R. Triggiani, Function space controllability of linear retarded systems: a derivation from abstract operator conditions, Report CRM-605, Centre de Recherches Mathématiques, Université de Montréal, 1976.

[9] A. Manitius and R. Triggiani, New results on functional controllability for time-delay systems, Proceedings of the 1976 Conference on Information Sciences and Systems, The Johns Hopkins University, Baltimore, Maryland 21218, pp. 401-405.

[10] A. Manitius and R. Triggiani, Sufficient conditions for function space controllability and feedback stabilizability of linear retarded systems, Proceedings of 1976 IEEE Conference on Decision and Control (to appear).

[11] A. Manitius, Controllability, observability and stabilizability of retarded systems, Proceedings of the 1976 IEEE Conference on Decision and Control (to appear).

[12] R. Triggiani, Controllability and observability in Banach space with bounded operators, SIAM J. Control 13 (1975), 462-491.

INDIRECT DELAY-FEEDBACK CONTROL OF LINEAR SYSTEMS

B. ASNER, JR.* AND A. HALANAY**
*Department of Mathematics, University of Dallas, Irving, TX
**University of Bucharest, Faculty of Mathematics, Bucharest, Romania

ABSTRACT

If the state and control spaces have the same dimensions then the state can be driven to zero in finite time by an indirect control law which is constructed using degenerate delay-differential equations. Discrete time control systems and on-line computer control are also constructed using past information of the state. Special boundedness and stability requirements are discussed.

INTRODUCTION

Consider the linear system

$$\dot{x} = Ax + u \tag{1}$$

where the system is fully controllable since the number of control parameters is equal to the number of state variables. We seek a control law u which stabilizes the system to

$$x(t) \equiv 0 \text{ for } t \geq t_1,\ t_1 \text{ finite.} \tag{2}$$

Adhering to the philosophy of Popov [7] that control laws consisting of linear delay terms are much simpler to construct than the corresponding nonlinear laws, we restrict our admissable feedback laws to be linear and to contain past information of the state. However, previous results in the theory of pointwise degenerate systems show that condition (2) is impossible to obtain if we restrict u to this class of feedback laws. In order to alleviate this situation, we propose a controller that is constructed from an auxiliary delay-system coupled with Eqn (1). Utilizing this type of indirect delay-feedback control and the theory of degenerate systems, we are able to achieve our primary goal, i.e., drive the full state of the system to zero in a finite time and keep it there. The structure of the feedback will then take the form

$$u(t) = \sum_{j=1}^{m}[B_j^x\, x(t-jh)+C_j^x\, \dot{x}(t-jh)+D_j^x z(t-jh)+E_j^x\, \dot{z}(t-jh)] \tag{3}$$

$$\dot{z}(t) = \sum_{j=0}^{m}[B_j^z x(t-jh)+C_j^z \dot{x}(t-jh)+ D_j^z z(t-jh)+E_j^z \dot{z}(t-jh)] \tag{4}$$

Several constructions which incorporate indirect delay-feedback are proposed and these may be categorized into two types: the constructions which incorporate instantaneous values in the auxiliary control law lead to systems with a better time value while those that do not utilize this information solve the problem, but with a sacrifice of time. In performing such constructions the general ideas in Kappel [6] have proved very useful.

All constructions are repeated for discrete-time control systems, Asner and Halanay [5]. The motivation for this is clearly seen if we think of the control device as being an on-line computer. Special stability requirements are discussed in this connection.

To fix notation, we let capital Latin letters be d x d matrices with complex entries, and lower case Latin letters be d-vectors. Some obvious exceptions: time is denoted by t and the delay time by h. The symbol x is reserved for the state of the system while the state of the "auxiliary" controller is denoted by z. Many of our proofs are simpler to obtain if we work with block matrices of higher order. For these cases we reserve the Greek alphabet and Italic script.

THE DELAY-FEEDBACK CONTROLLERS, CONTINUOUS CASE.

Table 1 contains control laws u, Eqn.3, and the corresponding auxiliary delay-systems, Eqn.4, for five distinct cases (A-E). In each case u(t) is obtained from delayed information on the state x and information on the auxiliary variable z (and y in Construction E). The main difference between Construction

---

Construction A: Three delays. $x(t)=0$ for $t\geq 3h$

$$u(t)=\frac{1}{h}[Bz(t-h)-2Be^{hA}z(t-2h)+Be^{2hA}z(t-3h)-e^{2hA}x(t-2h)+e^{3hA}x(t-3h)]$$

$$\dot{z}(t)=Az(t)+h^{-1}[e^{2hA}z(t-2h)-e^{3hA}z(t-3h)-B^{-1}e^{4hA}x(t-3h)].$$

Construction B: Two delays. $x(t) = 0$ for $t\geq 2h$.

$$u(t)=z(t)-2e^{hA}z(t-h)+e^{2hA}z(t-2h)-h^{-1}[e^{hA}x(t-h)-e^{2hA}x(t-2h)]$$

$$\dot{z}(t)=Az(t)+h^{-1}[e^{hA}z(t-h)-e^{2hA}z(t-2h)]-h^{-2}e^{2hA}x(t-2h).$$

Construction C: Delays using derivatives. $x(t)=0$ for $t\geq 2h$.

$$\begin{aligned}u(t)= h^{-1}B[(I+hA)z(t-h)-e^{hA}z(t-2h)]&+Ae^{hA}x(t-h)-h^{-1}e^{2hA}x(t-2h)\\ - B\dot{z}(t-h)\qquad\qquad &-e^{hA}\dot{x}(t-h)\end{aligned}$$

$$\dot{z}(t)=Az(t)-Ae^{hA}z(t-h)+e^{hA}\dot{z}(t-h)-B^{-1}e^{2hA}[Ax(t-h)-\dot{x}(t-h)] .$$

Construction D: Delays using derivatives. $x(t)=0$ for $t\geq h$.

$$u(t)=z(t)+(A-h^{-1}I)e^{hA}[x(t-h)+hz(t-h)]-e^{hA}[\dot{x}(t-h)+h\dot{z}(t-h)]$$

$$\dot{z}(t)=Az(t)- h^{-1}Ae^{hA}[x(t-h)+hz(t-h)]+h^{-1}e^{hA}[\dot{x}(t-h)+h\dot{z}(t-h)]$$

Construction E: One delay, two indirect controllers. $x(t)=0$ for $t\geq 2h$.

$$u(t)=y(t)-e^{hA}[2h^{-1}x(t-h)+3y(t-h)+hz(t-h)]$$

$$\dot{y}(t)=Ay(t)+z(t)+e^{hA}[2h^{-2}x(t-h)+4h^{-1}y(t-h)+3z(t-h)]$$

$$\dot{z}(t)=Az(t)-e^{hA}[2h^{-2}y(t-h)+2h^{-1}z(t-h)].$$

Restrictions: B is invertible, B commutes with A, and t, h>0. Constructions B, D, and E are special cases of more general situations which contain parameters.

TABLE 1. CONTINUOUS CASE

A and B is that in B we must have instantaneous information on z at time t in order to obtain u(t). This is not required in A where all terms used to construct u(t) are strictly delayed, but we must wait a longer time (2h vs. 3h) for the state to reach zero. Constructions C and D use delay-derivative control (neutral equations) and reduce the time for the state to reach zero by a factor of h. Construction C is similar to A as the control law u is obtained from past information, whereas D is similar to B. Construction E is a trade-off in time versus dimensionality. Information need only be retained for the past time interval h; however, the dimension of the auxiliary system increases to 2d and instantaneous information of one of the auxiliary variables is required.

We remark that Table 1 contains only special cases of more general situations. Some of the general cases are considered in the proofs of the Constructions.

Proof of Construction A: Let B be any matrix which is invertible and commutes with A, e.g., the identity matrix I. Introduce the 2d x 2d block matrices

$$\mathcal{A}=\begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix}, \quad \mathcal{B}_1=\begin{bmatrix} 0 & h^{-1}B \\ 0 & 0 \end{bmatrix}, \quad \mathcal{Z}=\begin{bmatrix} I \\ -B^{-1}e^{hA} \end{bmatrix} [e^{hA} \quad B]$$

the d x 2d block $Q^* = (I \quad 0)$, and define

$$\mathcal{B}_2=\mathcal{B}_1\mathcal{Z}-\mathcal{Z}\mathcal{B}_1 \qquad \mathcal{B}_3=-\mathcal{Z}\mathcal{B}_2 \ .$$

Define the 2d-vector $\chi=\begin{bmatrix} x \\ z \end{bmatrix}$ by the equation

$$\dot{\chi}(t) = \mathcal{A}\chi(t)+\mathcal{B}_1\chi(t-h)+\mathcal{B}_2\chi(t-2h)+\mathcal{B}_3\chi(t-2h), t>0 \tag{5}$$

and note that $\mathcal{Z}^2=0$ and $\mathcal{Z}\mathcal{A}=\mathcal{A}\mathcal{Z}$ since BA = AB.
For any solution of (5) define

$$\eta(t)=\chi(t)+\mathcal{Z}\chi(t-h). \tag{6}$$

Then for t≥h

$$\begin{aligned}
\dot{\eta}(t) &= \dot{\chi}(t)+\mathcal{Z}\dot{\chi}(t-h) \\
&= \mathcal{A}\chi(t)+\mathcal{B}_1\chi(t-h) + \mathcal{B}_2\chi(t-2h) +\mathcal{B}_3\chi(t-3h) \\
&\qquad +\mathcal{Z}\mathcal{A}\chi(t-h) +\mathcal{Z}\mathcal{B}_1\chi(t-2h) +\mathcal{Z}\mathcal{B}_2\chi(t-3h) +\mathcal{Z}\mathcal{B}_3\chi(t-3h) \\
&= \mathcal{A}\chi(t) +\mathcal{B}_1\chi(t-h)+\mathcal{A}\mathcal{Z}\chi(t-h) +\mathcal{B}_1\mathcal{Z}\chi(t-2h) \\
&= \mathcal{A}\eta(t)+\mathcal{B}_1\eta(t-h).
\end{aligned}$$

Thus,

$$\mathcal{B}_1\dot{\eta}(t) =\mathcal{B}_1\mathcal{A}\eta(t) = \mathcal{A}\mathcal{B}_1\eta(t) \quad \text{and} \quad \mathcal{B}_1\eta(t) = e^{\mathcal{A}t}c.$$

For t≥2h

$$\mathcal{B}_1\eta(t-h) = e^{\mathcal{A}(t-h)}c = e^{-h\mathcal{A}}\mathcal{B}_1\eta(t).$$

Substituting this the above equation for η(t) gives

$$\dot{\eta}(t) = (A+e^{-hA}B_1)\eta(t)$$

whose solution for $t \geq 3h$ can be expressed as

$$\eta(t) = e^{(A+e^{-hA}B_1)h}\eta(t-h) = \begin{bmatrix} e^{\mathrm{A}h} & \mathrm{B} \\ 0 & e^{\mathrm{A}h} \end{bmatrix} \eta(t-h) \qquad (7)$$

As $Q^* = (\mathrm{I}\ 0)$ we note that

$$Q^*Z=Q^*e^{(A+e^{-hA}B_1)h} \text{ and } Q^*e^{(A+e^{-hA}B_1)h}Z = 0 . \qquad (8)$$

Eqn (7) with (8) and (6) yield $x(t)=0$ for $t \geq 3h$.
$u(t)$ and the auxiliary controller of Table 1 are then obtained from the definitions of B_1, B_2, and B_3.

Proof of Construction B. Let matrices $\tilde{\mathrm{A}}$, $\hat{\mathrm{A}}$ be given such that the d x d matrix

$$\mathrm{D} = \int_0^h e^{\mathrm{A}(h-s)}\tilde{\mathrm{A}}e^{\hat{\mathrm{A}}s}ds$$

is invertible. Define the block matrices

$$A = \begin{bmatrix} \mathrm{A} & \tilde{\mathrm{A}} \\ 0 & \hat{\mathrm{A}} \end{bmatrix} \qquad Z = \begin{bmatrix} \mathrm{I} \\ -\mathrm{D}^{-1}e^{h\mathrm{A}} \end{bmatrix} [e^{h\mathrm{A}} \quad \mathrm{D}] \quad , \quad Q^* = (\mathrm{I} \quad 0),$$

and let

$$B_1 = AZ-ZA \text{ and } B_2 = -ZB_1 .$$

Using Eqn (5) with $B_3 = 0$ and the definition of Eqn (6), we obtain

$$\dot{\eta}(t) = A\eta(t) \text{ for } t \geq h$$

so that

$$\eta(t) = e^{hA}\eta(t-h) = \begin{bmatrix} e^{h\mathrm{A}} & \mathrm{D} \\ 0 & e^{h\hat{\mathrm{A}}} \end{bmatrix} \eta(t-h) \text{ for } t \geq 2h . \qquad (9)$$

Using the definition of η yields

$$\chi(t) + Z\chi(t-h)=e^{hA}[\chi(t-h) + Z\chi\ (t-h)] .$$

But $Q^*Z=Q^*e^{hA}$ and $Q^*e^{hA}Z = 0$.

The special case in Table 1 is obtained by letting $\tilde{\mathrm{A}}=\mathrm{I}, \hat{\mathrm{A}}=\mathrm{A}$, and the definitions of B_1, B_2. Note that $\mathrm{D}=he^{h\mathrm{A}}$ is invertible.

Proof of Construction C: Take χ, A, B_1 and Z as in Construction A. The relevant components of Table 1 are equivalent to

$$\dot{\chi}(t) = A\chi(t) + (B_1+AZ)\chi(t-h)+B_1Z\chi(t-2h) - Z\dot{\chi}(t-h).$$

Using (6) this last equation is equivalent to

$$\dot{\eta}(t)= A\eta(t) + B_1\eta(t-h),\ t \geq 0 .$$

Proceeding exactly as in the proof of Construction A yields for $t \geq 2h$ the desired result.

Proof of Construction D: Let $\tilde{\mathrm{A}}$, $\hat{\mathrm{A}}$, D, Z, and Q be defined as in Construction B, and consider the neutral equation

$$\dot{\chi}(t) = A\chi(t) + AZ\chi(t-h) - Z\dot{\chi}(t-h) \quad t>0 \tag{10}$$

which, utilizing Eqn (6), can be expressed as

$$\dot{\eta}(t) = A\eta(t).$$

Eqn (10) is valid for t>0 and the proof follows that of Construction B. Since

$$AZ = \begin{bmatrix} A-\tilde{A}D^{-1}e^{hA} \\ -\hat{A}D^{-1}e^{hA} \end{bmatrix} [e^{hA} \quad D] \qquad \text{and } Z = \begin{bmatrix} I \\ -D^{-1}e^{hA} \end{bmatrix} [e^{hA} \quad D]$$

we obtain the special case in Table I by letting $\tilde{A}=I$ and $\hat{A}=A$.

Proof of Construction E. Define $= \begin{bmatrix} A & I & 0 \\ 0 & A & I \\ 0 & 0 & A \end{bmatrix}$

$$\chi = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \qquad Q = \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix}, \qquad Z = \begin{bmatrix} I \\ -2h^{-1}I \\ 2h^{-2}I \end{bmatrix} \left(e^{hA} \quad he^{hA} \quad \frac{h^2}{2}\, e^{hA}\right)$$

$$B = AZ - ZA = \begin{bmatrix} -2h^{-1}e^{hA} & -3e^{hA} & -he^{hA} \\ 2h^{-2}e^{hA} & 4h^{-1}e^{hA} & 3e^{hA} \\ 0 & -2h^{-2}e^{hA} & -2h^{-1}e^{hA} \end{bmatrix}.$$

By direct computation we obtain $Z^2=0$, $ZAZ=0$ and this implies $ZB=0$. Use Eqn (6) for the system $\dot{\chi}(t)=A\chi(t)+B\chi(t-h)$ to obtain $\dot{\eta}(t)=A\eta(t)$ for $t\geq h$, and $\eta(t)=e^{hA}\eta(t-h)$ for $t\geq 2h$. The identities $Q^*e^{hA}Z=0$, $Q^*Z=Q^*e^{hA}$ shows that $x(t)\equiv 0$ for $t\geq 2h$.

DELAY-FEEDBACK CONTROLLERS, DISCRETE CASE.

In this section we consider the linear system

$$x_{n+1} = Ax_n + u_n\,, \quad n=0,\ 1,\ \ldots \tag{11}$$

where the state vector x_n and the control vector u_n are d-vectors. We also assume that A is invertible. As before, we seek a u_n to stabilize the system, i.e.,

$$x_n = 0 \quad \text{for } n\geq k\ . \tag{12}$$

The problem is trivial if no delays are allowed in the control law. In this case it suffieces to let $u_n = -Ax_n$ to obtain $x_n = 0$ for $n\geq 1$. In many problems, however, we wish to avoid this trivial solution as it requires an unbounded control law. For example, we may start with the continuous plant of Eqn (1) and use a sample-and-hold device to approximate u(t),

$$u(t)=u_n, \quad n\delta\leq t<(n+1)\delta\ . \tag{13}$$

Let $x(n\delta)=x_n$ to obtain the discrete system

$$x_{n+1} = e^{A\delta}x_n + B(\delta)\ u_n \qquad n=0,\ 1,\ \ldots \tag{14}$$

where

$$B(\delta) = \int_0^\delta e^{sA} ds . \tag{15}$$

Note that $B(\delta)$ is invertible for small δ but its inverse is of order δ^{-1}. Since the order of $e^{\delta A}$ is 1, a control law of the form

$$u_n = B^{-1}(\delta) e^{A\delta} x_n$$

is of order δ^{-1} and becomes unbounded as $\delta \to 0$. This situation may be avoided by including delay terms in u_n.

Table 2 is similar to Table 1 as we have attempted to parallel the five distinct cases. The structure of the complete system is of the form

$$\begin{aligned} x_{n+1} &= Ax_n + u_n \\ u_n &= f(x_n, \ldots, x_{n-\ell}, z_n, \ldots, z_{n-\ell}) \\ z_{n+1} &= g(x_n, \ldots, x_{n-\ell}, z_n, \ldots, z_{n-\ell}) \end{aligned}$$

where the functions f and g are assumed to be linear. The precise definitions of u_n, f, and g appear in Table 2:

Construction A': Three delays. $x_n = 0$ for $n \geq 3k$.

$$u_n = k^{-1}[Bz_{n-k} - 2BA^k z_{n-2k} + BA^{2k} z_{n-3k} - A^{2k+1} x_{n-2k} + A^{3k+1} x_{n-3k}]$$

$$z_{n+1} = Az_n + k^{-1}[A^{2k+1} z_{n-2k} - A^{3k+1} z_{n-3k} - B^{-1} A^{4k+2} x_{n-3k}]$$

Construction B': Two delays. $x_n = 0$ for $n \geq 2k$.

$$u_n = k^{-1}[z_n - 2A^k z_{n-k} + A^{2k} z_{n-2k} - A^{k+1} x_{n-k} + A^{2k+1} x_{n-2k}]$$

$$z_{n+1} = Az_n + k^{-1}[A^{k+1} z_{n-k} - A^{2k+1} z_{n-2k} - A^{2k+2} x_{n-2k}]$$

Construction C': Noncommensurate delays. $x_n = 0$ for $n \geq 2k$.

$$\begin{aligned} u_n = {} & B[(1+\tfrac{1}{k}) z_{n-k} - \tfrac{1}{k} A^k z_{n-2k} - A^{-1} z_{n+1-k}] \\ & + A^{k+1} x_{n-k} - \tfrac{1}{k} A^{2k+1} x_{n-2k} - A^k x_{n+1-k} \end{aligned}$$

$$z_{n+1} = Az_n - A^{k+1} z_{n-k} + A^k z_{n+1-k} - B^{-1} A^{2k+2} x_{n-k} + B^{-1} A^{2k+1} x_{n+1-k}$$

Construction D': Noncommensurate delays. $x=0$ for $n \geq k$.

$$u_n = \tilde{A} z_n + (AD - \tilde{A} D^{-1} A^k D) z_{n-k} - Dz_{n-k+1} + (A^{k+1} - \tilde{A} D^{-1} A^{2k}) x_{n-k} - A^k x_{n-k+1}$$

$$z_{n+1} = \hat{A} z_n - \hat{A} D^{-1} A^k D z_{n-k} + D^{-1} A^k D z_{n-k+1} - \hat{A} D^{-1} A^{2k} x_{n-k} + D^{-1} A^{2k} x_{n-k+1}$$

Construction E': One delay, two controllers. $x_n = 0$ for $n \geq 2k$.

$$u_n = k^{-1} y_n + (k+1)^{-1}[-2A^{k+1} x_{n-k} - \frac{(1+3k)}{k} A^k y_{n-k} - 2A^{k-1} z_{n-k}]$$

$$y_{n+1} = Ay_n + k^{-1} z_n + (k+1)^{-1}[2A^{k+2} x_{n-k} + 4A^k y_{n-k} + \frac{3k-1}{k} A^k z_{n-k}]$$

$$z_{n+1}=Az_n+(k+1)^{-1}[\qquad\qquad -2A^{k+2}y_{n-k}-2A^{k+1}z_{n-k}]$$

Restrictions: B is invertible, B commutes with A, n≥0. Constructions B and E are special cases of more general situations. The parameters in D are defined in the Proof of Construction B.

TABLE 2. DISCRETE CASES

Proof of Construction A': This proof is quite similar to that of Construction A.

Define

$$B_1=\begin{bmatrix}0 & k^{-1}B\\ 0 & 0\end{bmatrix},\quad Z=\begin{bmatrix}I\\ -B^{-1}A^{k+1}\end{bmatrix}(A^k\quad A^{-1}B),\quad \chi_n=\begin{bmatrix}x_n\\ z_n\end{bmatrix},$$

Q, B_2, and B_3 as in Construction A and consider the discrete system:

$$\chi_{n+1}=A\chi_n+B_1\chi_{n-k}+B_2\chi_{n-2k}+B_3\chi_{n-3k},\qquad n\geq 0\ .$$

Let $\eta_n=\chi_n+Z\chi_{n-k}$ (16)

and use the above relations to show that

$$\eta_{n+1}=A\eta_n+B_1\eta_{n-k}\ ,\ n\geq k\ .$$

Thus

$$B_1\eta_{n+1}=AB_1\eta_n\ ,\qquad B_1\eta_n=A^nc$$

and

$$\eta_{n+1}=(A+A^{-k}B_1)\eta_n\ ,\ n\geq 2k\ .$$

The solution to this last equation may be expressed as

$$\eta_n=(A+A^{-k}B_1)^k\eta_{n-k}=\begin{bmatrix}A^k & A^{-1}B\\ 0 & A^k\end{bmatrix}\eta_{n-k}\ ,\ n\geq 3k\ .$$

Since

$$Q^*(A+A^{-k}B_1)^k=Q^*Z \text{ and } Q^*Z^2=0 \text{ we have}$$

$$x_n=0 \text{ for } n\geq 3k\ .$$

Proof of Construction B': Define

$$D=\sum_{p=o}^{k-1}A^{k-p-1}\tilde{A}\,\hat{A}^p\ ;\qquad Z=\begin{bmatrix}I\\ -D^{-1}A^k\end{bmatrix}[A^k\quad D]$$

where D^{-1} is assumed to be invertible. A, B_1, B_2, and Q are defined as in Construction B with χ_n and η_n defined as in Construction A'. Thus we obtain

$$\chi_{n+1}=A\chi_n+B_1\chi_{n-k}+B_2\chi_{n-2k}\qquad n\geq 0$$

$$\eta_{n+1}=A\eta_n\qquad n\geq k$$

$$\eta_n=A^k\eta_{n-k}=\begin{bmatrix}A^k & D\\ 0 & A^k\end{bmatrix}\eta_{n-k},\ n\geq 2k,$$

and the result follows as before.

The special situation of Table 2 is obtained by letting $\tilde{A}=k^{-1}I$,

$\hat{A}=A$, and the definitions of B_1, B_2 . Note that $D=A^{k-1}$ is invertible under the assumption that A is invertible.

Proof of Construction C': Take χ, A, B_1, and Z as in Construction A'. The relevant components of Table 2 are equivalent to

$$\chi_{n+1} = A\chi_n+(B_1+AZ)+\chi_{n-k}+B_1Z\chi_{n-2k}-Z\chi_{n+1-k}, \quad n\geq 0 .$$

If we use Eqn (16), then this last equation is equivalent to

$$\eta_{n+1} = A\eta_n+B_1\eta_{n-k}, \quad n\geq 0 .$$

Proceeding exactly as in the proof of Construction A' yields, for $n\geq 2k$, the desired result.

Proof of Construction D': Let A, $\hat{A}$, $\tilde{A}$, D, Z , and Q be defined as in the proof of Construction B' and consider the equation

$$\chi_{n+1} = A\chi_n+AZ\chi_{n-k}-Z\chi_{n-k+1}, \quad n\geq 0 .$$

The standard substitution of Eqn (16) reduces this to

$$\eta_{n+1} = A\eta_n, \quad n\geq 0 .$$

Thus

$$\eta_n = A^k\eta_{n-k} = \begin{bmatrix} A^k & D \\ 0 & A^k \end{bmatrix} \eta_{n-k}, \quad n\geq k,$$

and the result follows when we multiply this last expression by Q*.

Proof of Construction E :

$$\text{Define } A = \begin{bmatrix} A & k^{-1}I & 0 \\ 0 & A & k^{-1}I \\ 0 & 0 & A \end{bmatrix}, \quad \chi = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \quad Q = \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix},$$

$$Z = \begin{bmatrix} I \\ -\frac{2k}{k+1}A \\ \frac{2k}{k+1}A^2 \end{bmatrix} \begin{bmatrix} A^k & A^{k-1} & \frac{k-1}{2k}A^{k-2} \end{bmatrix},$$

$$B = AZ - ZA = \frac{1}{k+1} \begin{bmatrix} -2A^{k+1} & -\frac{1+3k}{k}A^k & -2A^{k-1} \\ 2A^{k+2} & 4A^{k+1} & \frac{3k-1}{k}A^k \\ 0 & -2A^{k+2} & -2A^{k+1} \end{bmatrix},$$

and write the system as

$$\chi_{n+1} = A\chi_n + B\chi_{n-k}, \quad n\geq 0 .$$

Direct computation yields $Z^2=ZAZ=0$ and $ZB=0$. With η_n defined as above, we obtain $\eta_{n+1}=A\eta_n$ for $n\geq k$, and $\eta_n=A^k\eta_{n-k}$ for $n\geq 2k$. Now use the definition of η_n with the identities $Q^*A^kZ=0$ and $Q^*Z=Q^*A^k$ to show that $x_n=0$ for $n\geq 2k$.

STABILITY AND BOUNDEDNESS

As we remarked earlier in this paper, a trivial control law with no delays can lead to an unbounded control law while the proper utilization of delay terms alleviates this undesired phenomenon.

For convenience we repeat the pertinent ideas as stated earlier. That is, given the continuous plant

$$\dot{x}(t) = Ax + u(t),$$

approximate

$$u(t) = u_n \ , \ n\delta \le t < (n+1)\delta,$$

and let

$$x(n\delta) = x_n$$

to obtain the discrete system

$$x_{n+1} = e^{\delta A} x_n + v_n \qquad n = 0,1,\ldots \tag{17}$$

where

$$v_n = \int_0^\delta e^{sA} ds \ \ u_n \ . \tag{18}$$

We now choose v_n to be any of the control laws defined in Table 2 by formally substituting $e^{\delta A}$ for A and v_n for u_n. The parameter k is defined by

$$k = [\tfrac{h}{\delta}]$$

where [ ] is the greatest integer function and h is the apriori basic delay interval associated with the continuous plant.

For any of the constructions,

$$x_n = x_{n+1} = 0 \text{ after } n \ge 3k \text{ and from Eqn (17), } v_n = 0.$$

This implies, from Eqn (18) and δ fixed that $u_n=0$, $n \ge 3k$. (19)

Referring again to Table 2, we observe that

$$(e^{\delta A})^k = e^{\delta k A} \approx e^{hA}$$

remains constant, while $k^{-1} \to 0$ as $\delta \to 0$. Hence, an apriori bound on the auxiliary controllers will imply an ϵ-bound for v_n in Constructions A', B', and E'. This apriori bound can be satisfied, for example, if A is stable.

The behavior of x(t) in the time interval $n\delta \le t < (n+1)\delta$ satisfies

$$x(t) = e^{A(t-n\delta)} x(n\delta) + \int_{n\delta}^{t} e^{A(t-s)} u_n ds \ .$$

For fixed δ and $n \ge 3k$, $x(n\delta) = x_n = 0$, and from Eqn (19), $u_n = 0$. Hence, x(t) eventually becomes and remains zero. That is,

$$x(t) \equiv 0 \text{ for } t \ge n\delta \ge 3k\delta \approx 3h.$$

This bound is lowered to 2h for Construction B', C', and E', and to h for Construction D'.

REFERENCES

[1] Asner, B.A., "New Constructions for Pointwise Degenerate Systems," IFAC, August, 1975.

[2] Asner, B.A. and A. Halanay, "Delay-Feedback Solution for the Minimal-Time Linear Problem with Several Outputs," REVUE ROMAINE DE MATHEMATIQUES PURES ET APPLIQUEES, Tome XVIII, No. 2, Bucharest (1973), 295-6.

[3] ____________, "Delay-Feedback Using Derivatives for Minimal Time Linear Control Problems," J.MATH. ANAL. AND APPLICATIONS, Vol. 48, No. 1, October 1974.

[4] ____________, "Pointwise Degenerate Second Order Delay-Differential Systems," ANALELE UNIVERSITATII BUCURESTI, MATEMATICA-MECANICA, Anul.XXII, No. 2, 1973.

[5] ____________, "Minimal-Time Delay-Feedback for Discrete Systems," REVUE ROUMAINE DES SCIENCES TECHNIQUES, Tome 20, Bucharest (1975).

[6] Kappel, F. "Degenerate-Differential Equations" Algebraic Theory," to appear.

[7] Popov, V.M., "Delay-Feedback, Time-Optimal, Linear Time-Invariant Control Systems," "Ordinary Differential Equations, 1971 NRL-MRC Conference," Academic Press, New York (1972).

COMMON AND PRIVATE INFORMATION BETWEEN TWO PERSONS

Y. C. Ho
Division of Engineering and Applied Physics
Harvard University
Cambridge, MA 02138

ABSTRACT

This is an informal talk based on work with G. Hexner and C. Martin which appears in [1]. Briefly, we are interested in making precise the colloquial notions of "What you and I both know" and "What I know that you don't know" from a decision-theoretic viewpoint. It seems that such questions are basic in any study of large scale multi-person information systems, such as, command-control systems, and database systems.

We propose certain definitions to characterize the common and private information between two persons and show that they satisfy intuitively reasonable properties.

REFERENCE

[1] Y. C. Ho, G. Hexner, and C. Martin, "Two Person Information Structure", Division of Engineering and Applied Physics, Harvard University, Technical Report No. 663, July 1976; portion to appear in the IEEE Transactions on Information Theory and the International Journal on Control.

DECENTRALIZED ROBUST CONTROL OF UNKNOWN SYSTEMS USING TUNING REGULATORS

E.J. DAVISON

Dept. of Electrical Engineering
University of Toronto
Toronto, Canada M5S 1A4

ABSTRACT

The problem of determining a decentralized robust controller for a general servomechanism problem, in which the system to be regulated is unknown, is considered in this paper. The only assumptions made regarding the system to be regulated are that (i) the system is described by a finite linear time-invariant multivariable model, (ii) the system is stable. Note that the order of the system is not assumed to be known. The approach taken to solve the problem is in the same spirit of some results recently obtained on "multivariable tuning regulators" for centralized control systems [1]. It is shown that necessary and sufficient conditions for a solution to exist to the problem can be obtained by performing some simple experiments on the plant, and that a robust decentralized controller can then be synthesized using the results of these experiments in conjunction with a series of one-dimensional parameter searches performed on the closed loop system. This later procedure is called "tuning the controller on line". The most interesting feature of this problem which differs from the centralized control problem is that the controller synthesis must be carried out in a certain preassigned order with respect to each controller agent (not necessarily unique); if this ordering is not carried out, then there will exist no controller synthesis for the problem.

1. INTRODUCTION

In [1], the following type of problem was considered: given an unknown system which is assumed to be stable, is it possible to find a centralized robust controller for the system so that closed loop stability and asymptotic regulation occur independent of any disturbances occurring in the system. In this paper, the same problem as treated in [1] will be considered, except that the controller structure will now be constrained to be decentralized [2]. The motivation for studying this problem is that such a decentralized problem is extremely important when studying large scale system problems, e.g. electric power systems [3], socioeconomic systems [4], chemical process control [5] etc.

2. DEVELOPMENT

The following problem is considered:

2.1 Problem Statement

Consider the following linear, time-invariant, stable system (plant) consisting of ν control agents:

This work has been supported by the National Research Council of Canada under Grant No. A4396.

$$\dot{x} = Ax + \sum_{i=1}^{\nu} B_i u_i + E\omega$$

$$e_i = C_i x + D_i u_i + F_i \omega - y_i^{ref}, \quad i = 1, 2, \ldots, \nu \tag{1}$$

where $x \in R^n$ is the state, $u_i \in R^{m_i}$ is the i^{th} control agent's control input, $e_i \in R^{r_i}$ is the i^{th} control agent's error which is the difference between the i^{th} control agent's output $y_i \triangleq C_i x + D_i u_i + F_i \omega$ and the specified i^{th} control agent's reference input y_i^{ref}. $\omega \in R^q$ is an input disturbance.

Let

$$y_{ref} \triangleq \begin{bmatrix} y_1^{ref} \\ \vdots \\ y_\nu^{ref} \end{bmatrix}, \quad e \triangleq \begin{bmatrix} e_1 \\ \vdots \\ e_\nu \end{bmatrix}, \quad u \triangleq \begin{bmatrix} u_1 \\ \vdots \\ u_\nu \end{bmatrix}, \quad E \triangleq \begin{bmatrix} E_1 \\ \vdots \\ E_\nu \end{bmatrix}, \quad F \triangleq \begin{bmatrix} F_1 \\ \vdots \\ F_\nu \end{bmatrix}, \quad C \triangleq \begin{bmatrix} C_1 \\ \vdots \\ C_\nu \end{bmatrix},$$

$$B \triangleq (B_1, B_2, \ldots, B_\nu), \quad D \triangleq \text{block diag } (D_1, D_2, \ldots, D_\nu) \tag{2}$$

In this problem description, it is assumed that the order n and the plant parameters A, B, C, D, E, F are unknown. In addition, it is not necessarily assumed that rank $B_i = m_i$, rank $C_i = r_i$. It is assumed however that A is stable, that the control inputs u_i can be excited, that the outputs y_i can be measured and that y_i^{ref} is known. For simplicity, it is assumed in the development to follow that $m_i = r_i$, $i = 1, 2, \ldots, \nu$.

It is assumed that the disturbance vector ω satisfies the equation:

$$\dot{\eta}_1 = A_1 \eta_1, \quad \eta_1 \in R^{n_1}$$

$$\omega = C_1 \eta_1 \tag{3}$$

where (C_1, A_1) is observable, and that the specified reference input y_{ref} satisfies the equation:

$$\dot{\eta}_2 = A_2 \eta_2, \quad \eta_2 \in R^{n_2}$$

$$\rho = C_2 \eta_2$$

$$y_{ref} = G\,\rho \tag{4}$$

where (C_2, A_2) is observable. For nontrivality, it is assumed that $\text{rank}\begin{bmatrix} E \\ F \end{bmatrix} = \text{rank } C_1$, rank G = rank C_2. Let the minimal polynomial of A_1, A_2 be denoted by $\Lambda_1(s)$, $\Lambda_2(s)$ respectively, and let the zeros of the least common multiple of $\Lambda_1(s)$, $\Lambda_2(s)$ (multiplicies included) be given by λ_k, $k = 1, 2, \ldots, \bar{p}$, defined so that λ_k is repeated q_k times, $k = 1, 2, \ldots, \bar{p}$, where λ_k, $k = 1, 2, \ldots, \bar{p}$ are all distinct from each other; then, it is assumed that $\text{Re}(\lambda_k) = 0$, $k = 1, 2, \ldots \bar{p}$ and that λ_k, q_k, $k = 1, 2, \ldots \bar{p}$ are known. This corresponds to the case of finding a robust decentralized controller for (1) for the case of a specified class of polynomial-sinusoidal disturbance/tracking signals. Note that it is not assumed that $(C_1, A_1; C_2, A_2, G)$ is necessarily known.

The following definitions will be used in the development to follow:

Definitions

Given $M = (m_{ij})$; let $M \in \Omega_\epsilon$ where $\Omega_\epsilon = \{x | |x| < \epsilon\}$.

Decentralized Robust Control Problem

It is desired now to final a decentralized controller [2] for (1) so that the following conditions are all satisfied:

(i) The closed loop system is stable, i.e. $\lim_{t\to\infty} \dot{x}(t) = 0$, $\forall x(0) \in R^n$, with $\eta_1(0) = 0$, $\eta_2(0) = 0$.

(ii) Asymptotic regulation takes place, i.e. $\lim_{t\to\infty} e(t) = 0$, $\forall x(0) \in R^n$, $\forall \eta_1(0) \in R^{n_1}$, $\forall \eta_2(0) \in R^{n_2}$.

(iii) If the plant parameters A, B, C are perturbed, i.e. $A \to A + \delta A$, $B \to B + \delta B$, $C \to C + \delta C$ and if $\epsilon > 0$ is chosen so that the resultant closed loop system remains stable $\forall \delta A \in \Omega_\epsilon$, $\forall \delta B \in \Omega_\epsilon$, $\forall \delta C \in \Omega_\epsilon$, it is desired that asymptotic regulation still occur for the resultant perturbed system.

In this problem formulation, it is assumed that the outputs y_i, y_i^{ref}, of the i^{th} control agent may only be used as inputs to the i^{th} control agent's controller (which has u_i for its output). This constraint arises from the decentralized formulation imposed in the problem. A controller synthesis for the problem then consists of finding ν separate controllers for the ν control agents to satisfy the three requirements listed above.

Remark 1

In this problem formulation, it is assumed that the outputs y_i are available for measurement. It should be noted that no loss of generality occurs by making this assumption, since it is shown in [2] that this is a necessary condition for there to exist a solution to the problem.

2.2 Method of Solution

Since the plant is unknown, the problem may be considered to be a compensator identification problem and following [1], it is desired to know if there exists a finite set of experiments (taking into account noisy measurements) to perform on the plant, so that one can express the necessary and sufficient conditions for there to exist a solution to the above problem in terms of these experiments. Assuming that a solution to the problem exists, it is then desired to know if there exists a controller synthesis for the problem, which satisfies the following decentralized controller tuning synthesis constraint.

2.3 Decentralized Controller Tuning Synthesis Constraint

In obtaining a controller synthesis for the plant, the following constraints are imposed:

(i) A controller can be implemented on only one control agent at any given time.

(ii) After a controller has been obtained and implemented on a given control agent, it is assumed that this controller is fixed and cannot be readjusted.

(iii) At all times during the controller synthesis, the resultant closed loop system must remain stable.

Remark 2

The motivation for imposing condition (iii) is obvious. The motivation for imposing conditions (i), (ii) is that such conditions correspond to a large class of practical problems in large scale systems, e.g. process control.

The following definitions will be used in the development to follow:

Definition

If there exists a decentralized controller synthesis which satisfies the decentralized controller tuning synthesis constraint and which solves the decentralized robust control problem, then it is said that there exists a solution to the robust decentralized servomechanism problem for unknown systems.

Definition

The tracking steady-state gain parameters $T_k(i,j)$, $i = 1, 2, \ldots, \nu$; $j = 1, 2, \ldots,$; $k = 1, 2, \ldots, \bar{p}$ of the system (1) are defined to be as follows:

$$T_k(i,j) \triangleq C_j(\lambda_k I - A)^{-1}B_i + D_j \qquad (5)$$

3. SOLUTION OF THE PROBLEM

The following main result is obtained:

Theorem 1

There exists a solution to the robust decentralized servomechanism problem for unknown systems if and only if there exist of list of integers $(s_1, s_2, \ldots, s_\nu)$ (not necessarily unique) where $s_i \in [1, 2, \ldots, \nu]$ and s_i, $i = 1, 2, \ldots, \nu$ are all distinct, such that the following ν conditions all hold:

$$(1) \quad \text{rank}[T_k(s_1,s_1)] = r_{s_1}, \; k = 1,2,\ldots,\bar{p}$$

$$(2) \quad \text{rank}\begin{bmatrix} T_k(s_2,s_2), & T_k(s_2,s_1) \\ T_k(s_1,s_2), & T_k(s_1,s_1) \end{bmatrix} = r_{s_1} + r_{s_2}, \; k = 1,2,\ldots\bar{p}$$

$$(3)\quad \operatorname{rank}\begin{bmatrix} T_k(s_3,s_3), & T_k(s_3,s_2), & T_k(s_3,s_1) \\ T_k(s_2,s_3), & T_k(s_2,s_2), & T_k(s_2,s_1) \\ T_k(s_1,s_3), & T_k(s_1,s_2), & T_k(s_1,s_1) \end{bmatrix} = r_{s_1} + r_{s_2} + r_{s_3}, \quad k = 1, 2, \ldots, \bar{p}$$

$$\vdots$$

$$(\nu)\quad \operatorname{rank}\begin{bmatrix} T_k(s_\nu,s_\nu), & T_k(s_\nu,s_{\nu-1}), \ldots, & T_k(s_\nu,s_1) \\ \vdots & \vdots & \vdots \\ T_k(s_1,s_\nu), & T_k(s_1,s_{\nu-1}), \ldots, & T_k(s_1,s_1) \end{bmatrix} = \sum_{i=1}^{\nu} r_i, \quad k = 1, 2, \ldots \bar{p} \qquad (6)$$

Proof

Due to space limitations, all proofs are omitted; they may be found in [6].

Discussion

It can be seen from theorem 1, that the existence of a solution to the robust decentralized servomechanism problem for unknown systems can be determined if the steady-state gain parameters of the system are known. It will now be shown that these gain parameters can easily be calculated by performing a set of experiments on (1).

3.1 Experimental Calculation of Steady-state Tracking Parameters

Let $y \triangleq \begin{bmatrix} y_1 \\ \vdots \\ y_\nu \end{bmatrix}$; then the steady-state tracking gain parameters of the system may be found by using exactly the same steady-state experiments on (1) as outlined in [1] for the centralized case; in this case, the input u_i is excited with $u_j = 0$, $j = 1, 2, \ldots, \nu$, $j \neq i$ and the output y is measured. In these experiments, it is only necessary to measure the steady-state ouput of the system. This can usually be done quite accurately in the prescence of a large class of noise processes (unlike the case when measuring the transient outputs of the system).

3.2 Controller Synthesis

Assume that theorem 1 holds; then a controller synthesis for the decentralized robust servomechanism problem for unknown systems proceeds as follows:

Step 1. Apply a centralized servo-compensator [1] to control agent s_1 using the centralized tuning algorithm given in [1] for the class of disturbance/tracking signals described by (3), (4). This is always possible to do if condition 1 of theorem 1 holds. In this case, the controller for agent s_1 is found by applying a number of on-line tuning procedures (one-dimensional searches) so that "maximum speed of response" is obtained.

Step 2. Find the steady-state parameter $T_k(s_2,s_2)$, $k = 1, 2, \ldots, \bar{p}$ for the resultant closed loop augmented system obtained above, by performing a set of experiments on the system as discussed in section 3.1. Then, using these steady-state paramters, apply a centralized servo-compensator to control agent s_2 using the centralized tuning algorithm given in [1]. This is always possible to do if conditions 1, 2 of theorem 1 hold.

⋮

Step ν. Find the steady-state paramters $T_k(s_\nu, s_\nu)$, $k = 1, 2, \ldots, \bar{p}$ for the resultant closed loop augmented system obtained in step ν-1; then using these steady-state paramters, apply a centralized servo-compensator to control agent s_ν using the centralized tuning algorithm given in [1]. This is always possible to do if condition ν of theorem 1 holds.

Remark 3

All ν steps in the above controller synthesis can always be carried out so that the resultant closed loop system remains stable.

4. DISCUSSION OF PROBLEM SOLUTION

It is interesting to compare the conditions for a solution of the present problem, given by theorem 1, with those required when the decentralized constraint is removed. It is shown in [1] that for the centralized control of unknown systems, the corresponding necessary and sufficient condition is given by condition ν of theorem 1. Thus it is seen that when the decentralization constraint is imposed, the condition required for a solution to exist becomes more demanding than when the decentralization constraint is removed. It should be noted however that if $m_i \geq r_i$, $i = 1, 2, \ldots, \nu$ the conditions of theorem 1 will "almost always" hold, i.e. the class of (C,A,B,D) systems in which theorem 1 does not hold lies on a hypersurface in the parameter space of (C,A,B,D).

It is also interesting to note the main difference occurring between the controller synthesis for the present problem, and for the case when the decentralization constraint is removed. In the case of centralized control, the controller is found in one step (via a number of tuning procedures [1]), whereas in the case of decentralized control, the controller synthesis must be carried out in a certain preassigned order (not necessarily unique) with respect to each controller agent; if the controller synthesis is not carried out in this special order, then in general, no controller synthesis can be obtained.

The necessity of carrying out the tuning of controller agents in a particular sequence is interesting and possibly gives some insight into some practical problems of process control in which decentralized controllers are often used in practice (the controllers being found heurestically), i.e. it explains possibly why it sometimes occurs in two identical industrial processes that in one case immense difficulties are found in controlling the system, whereas in another case no difficulties at all are experienced in controlling the system. The possible explanation is that in the two cases, the order of applying decentralized controllers (via tuning methods) is not necessarily the same, and that in the first case, the incorrect order is being used.

The necessity of tuning controller agents in a particular order also gives possible insight into social-economic systems and other large scale system problems, e.g. when decentralized heurestic controllers are attempted to be found to control certain social-economic systems, it may well be that no solution exists unless the decentralized controllers are applied in a particular order.

5. NUMERICAL EXAMPLES

The following example is given to illustrate the previous results. Consider the system:

$$\dot{x} = Ax + B_1u_1 + B_2u_2 + E\omega \ , \quad e_1 = y_1 - y_1^{ref}$$
$$y_1 = C_1x + D_1u_1 + F_1\omega_1 \ , \quad e_2 = y_2 - y_2^{ref}$$
$$y_2 = C_2x + D_2u_2 + F_2\omega_2 \ , \tag{7}$$

where

$$A = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad C_1 = (1 \quad 0),$$

$$C_2 = (0 \quad 1), \ E = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \ F_1 = 0, \ F_2 = 0, \ D_1 = 0, \ D_2 = 0$$

and ω is a constant disturbance. It is desired to find a decentralized robust controller for (7) so that the outputs asymptotically are equal to certain preassigned set points y_1^{ref}, y_2^{ref}. The order of the system and matrices A, B_1, B_2, C_1, C_2, E, F_1, F_2, D_1, D_2 are unknown. In this case $\bar{p} = 1$ and $\lambda_k = 0$, $k = 1$, and $\nu = 2$.

The following experiments, following section 3.1 and [1], are now carried out.

Experiment

$$\text{Let } u_1 = 1, \ u_2 = 0, \ t \geq 0; \text{ then } \lim_{t\to\infty} \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

$$\text{let } u_2 = 1, \ u_1 = 0, \ t \geq 0; \text{ then } \lim_{t\to\infty} \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \tag{8}$$

so that $T_1(1,1) = 0$, $T_1(1,2) = -1$, $T_1(2,1) = -1$, $T_1(2,2) = -1$. On noting now that:

$$\text{(i)} \quad \text{rank}[T_1(2,2)] = 1$$

$$\text{(ii)} \quad \text{rank} \begin{bmatrix} T_1(1,1), & T_1(1,2) \\ T_1(2,1), & T_1(2,2) \end{bmatrix} = 2 \tag{9}$$

it is seen that theorem 1 holds. The following controller synthesis is now carried out.

Step 1

Apply the controller [1]

$$\dot{\eta}_2 = 0\ \eta_2 + (y_2 - y_2^{ref})$$

$$u_2 = \epsilon_2[T_1(2,2)]^{\dagger}\eta_2,\ \epsilon_2 > 0 \qquad (10)$$

to agent 2, where ϵ_2 is found by performing a one-dimensional on-line search on the resultant closed loop system so that "maximum speed of response occurs"; in this case $T_1(2,2) = -1$ so that (10) becomes $u_2 = -\epsilon_2\eta_2$. Table 1 shows how the eigenvalues of the resultant closed loop system behaves as a function of ϵ_2.

Table 1 Variation of Closed-loop Eigenvalues with Respect to ϵ_2

ϵ_2	Eigenvalues		
0.05	-0.053	-0.95	-1.0
0.10	-0.11	-0.89	-1.0
0.15	-0.18	-0.82	-1.0
0.20	-0.28	-0.72	-1.0
0.25*	-0.50	-0.50	-1.0
0.30	-0.50 ± j0.22		-1.0
0.50	-0.50 ± j0.50		-1.0
1.00	-0.50 ± j0.87		-1.0

In this case, it can be seen that as ϵ_2 increases from zero, the system begins to oscillate; this corresponds very closely to the classical procedure of "tuning" a controller on line. A choice of $\epsilon_2 = 0.25$ is acceptable.

Step 2

Assuming that $\epsilon_2 = 0.25$ has been chosen for controller agent 2, the following experiment is now performed on the resultant closed loop system obtained by augmenting (10) with (7):

Let $u_1 = 1$, $u_2 = 0$, $t \geq 0$; then $\lim_{t\to\infty} y_1(t) = 1$, which means that $T_1(1,1)$ for the resultant controlled system is given by $T_1(1,1) = 1$. Now apply the following controller to the above controlled system:

$$\dot{\eta}_1 = 0\ \eta_1 + (y_1 - y_1^{ref})$$

$$u_1 = \epsilon_1\{1\}^{\dagger}\eta_1,\ \epsilon_1 > 0$$

$$= \epsilon_1\eta_1 \qquad (11)$$

where ϵ_1 is found by performing a one-dimensional search on the resultant closed loop system so that "maximum speed of response occurs". Table 2 shows how the eigenvalues of the resultant closed loop system behaves as a function of ϵ_1.

Table 2 Variation of Closed-loop Eigenvalues with Respect to ϵ_1

ϵ_1	Eigenvalues			
0.01	-0.011	-0.40	-0.60	-0.99
0.05	-0.075	-0.24	-0.76	-0.93
0.10*	-0.12 ± j0.13		-0.88 ± j0.13	
0.15	-0.10 ± j0.19		-0.90 ± j0.19	
0.20	-0.083 ± j0.22		-0.92 ± j0.22	
0.30	-0.053 ± j0.27		-0.95 ± j0.27	
0.50	-0.011 ± j0.34		-0.99 ± j0.34	
1.00	+0.059 ± j0.43		-1.06 ± j0.43	

In this case, it can be seen that as ϵ_1 increases from zero, the system begins to oscillate and eventually becomes unstable. A choice of $\epsilon_1 = 0.10$ is acceptable.

The final decentralized robust controller obtained for the system (7) is now given as follows:

$$\dot{\eta}_1 = 0\,\eta_1 + (y_1 - y_1^{ref}),\quad u_1 = 0.10\,\eta_1$$

$$\dot{\eta}_2 = 0\,\eta_2 + (y_2 - y_2^{ref}),\quad u_2 = 0.25\,\eta_2 \tag{12}$$

Remark 4

It should be noted that in this problem, there exists no controller synthesis if a controller is first implemented for control agent 1. In order to provide a solution, any controller synthesis must begin with control agent 2.

6. CONCLUSIONS

A study of the decentralized servomechanism problem for an unknown system is made in this paper. Necessary and sufficient conditions for a controller to exist for the problem and a controller synthesis procedure are given. A numerical example is given to illustrate the procedure.

REFERENCES

1. Davison, E.J., "Multivariable Tuning Regulators: The Feedforward and Robust Control of a General Servomechanism Problem", IEEE Trans. on Aut. Control, Vol. AC-21, No. 1, 1976, pp. 35-47.

2. Davison, E.J., "The Robust Decentralized Control of a General Servomechanism Problem", IEEE Trans. on Aut. Control, Vol. AC-21, No. 1, 1976, pp. 14-24.

3. Davison, E.J., Rau, N.S., Palmay, F.V., "The Optimal Decentralized Control of a Power System Consisting of a Number of Interconnected Synchronous Machines", Int. J. Control, Vol. 18, No. 6, 1973, pp. 1313-1328.

4. Varaiya, P., "Trends in the Theory of Decision-making in Large Systems", Ann. Econ. Soc. Meas., Vol. 1/4, 1972, pp. 493-500.

5. Foss, A.S., "Critique of Chemical Process Control Theory", IEEE Trans. on Automatic Control, Vol. AC-18, 1973, pp. 646-652.

6. Davison, E.J., "Decentralized Robust Control of Unknown Systems using Tuning Regulators", Dept. of Electrical Engineering, University of Toronto, System Control Report No. 7617 , October 1976.

ON STOCHASTIC GAMES WITH NONCOMPACT STRATEGY SPACES AND LOSS FUNCTIONS THAT EXHIBIT A PRODUCT STRUCTURE: WITH APPLICATION TO LEG STOCHASTIC DIFFERENTIAL GAMES

MARY ELLEN BOCK
Department of Statistics
Purdue University
West Lafayette, Indiana

MAX MINTZ
Department of Systems Engineering
Moore School
University of Pennsylvania
Philadelphia, Pennsylvania

1. Introduction:

We study existence theorems for saddle-point solutions for the following general class of statistical games:

Let $\mathcal{F} \overset{\Delta}{=} \{F_\theta(\cdot): \theta \epsilon R^n\}$ denote a family of probability distribution functions indexed by the unknown vector θ. Let Z denote an observation of a random vector with distribution function in $\mathcal{F}$. Suppose a Decision-Maker wishes to estimate the true value of θ in R^n given the observation Z. In this context, let $L(\delta,\theta)$ denote the Decision-Maker's loss function.

1.1 General Problem Statement:

Our objective will be to determine sets of sufficient conditions on $L(\delta,\theta)$ and $\mathcal{F}$ which insure that this statistical decision problem admits a least favorable prior distribution λ^* for θ and a minimax decision rule δ^* which is Bayes with respect to λ^*.

In order to solve this problem we must address some challenging questions. In particular, since R^n is not compact in the usual topology, the existence of (i) a least favorable prior distribution for θ, and (ii) a minimax Bayes rule, is by no means assured. For example, we cite the following well-known results [Ferguson, 1967; Strawderman, 1971]:

If $Z \sim N[\theta,\Sigma]$, and $L(\delta,\theta)$ is a standard quadratic loss function, then

(i) There is no least favorable prior distribution for $\theta \epsilon R^n$.

(ii) If $n \leq 2$, there is a minimax admissible rule which is generalized Bayes and unique up to equivalence.

(iii) If $n > 2$, a unique solution does not exist, i.e., there are both Bayes minimax rules, and non-Bayes minimax rules.

In game theory, the determination of saddle-point results is very often based on (i) "continuity" and "convexity" properties associated with the pay-off function, and (ii) the topology and convexity of the underlying strategy spaces. [See for example, Sion, 1958].

For our given decision problem, let $\mathcal{S}$ denote the class of all decision rules δ, and $\mathcal{M}$ the class of all probability measures on $(R^n, \mathcal{B}^n)$. In the following analysis we will determine suitable characteristics for $L(\delta,\theta)$ and $\mathcal{F}$, and suitable topologies on $\mathcal{S}$ and $\mathcal{M}$ which will guarantee that (i) all the "good strategies" for each player (the Decision-Maker vs. Nature) must lie in a compact convex subset of each player's respective strategy space, and (ii) the kernel of the game is suitably continuous and convex.

1.2 Preliminary Discussion:

As a preview to a detailed discussion, we outline briefly some of our underlying prior assumptions, and some of the salient features of our results.

(1). We focus our attention on a location parameter problem, i.e., let F denote a distribution function on R^n, then $\mathcal{F} = \{F_\theta(Z) \triangleq F(Z-\theta):\theta\varepsilon R^n\}$.

(2). Our underlying loss function $L(\delta,\theta)$ will exhibit the following product structure:

$$L(\delta,\theta) = L_1(\delta-\theta)L_2(\theta) ,$$

where L_1: $R^n \to R$ is a continuous convex function which is bounded below and L_2: $R^n \to R$ is a nonnegative continuous function satisfying suitable growth conditions.

The use of a loss function from this class would be reasonable, when the Decision-Maker is confronted by an intelligent adversary (Nature) who is faced with a trade-off between a gain due to estimation error [the factor $L_1(\delta-\theta)$] and a loss due to a choice of θ with large $||\theta||$ [the factor $L_2(\theta)$]. Loss functions in this class are useful in modelling the decision problem associated with tracking on evasive target, where we assume the evasive target is under the control of an intelligent adversary, who must balance a gain due to estimation error against a loss due to energy consumed in evasive maneuvers. In a recent paper [Stelte and Mintz, 1973] it has been shown that if

(i) $L(\delta,\theta) = e^{c(\delta-\theta)^2}e^{-k\theta^2}$, where $c \geq 0$ and $k > 0$ are given scalar weighting coefficients, and

(ii) $Z \sim N[\theta, \sigma^2]$, $\theta\varepsilon R$, where $c < k + \frac{1}{2\sigma^2}$, then there is a minimax estimate δ^*, and a least favorable prior distribution λ^*. Further, δ^* is linear and λ^* is normal.

The purpose of the present paper is to determine existence results for (i) a saddle-point theorem, (ii) a least favorable prior, and (iii) a Bayes minimax rule, for a very large class of statistical games of which the previous result is just a special case.

(3). In determining the existence of our saddle-point result, and the existence of a least favorable prior distribution, we make use of the weak topology on the space of distribution functions. In particular, using the characteristics of L_1 and L_2, and the direct half of Prohorov's theorem, we obtain a weakly compact set which contains all least favorable distributions. This weakly compact set is also used in the application of a saddle-point result due to Sion, which plays an essential role in obtaining our final saddle-point result.

(4). In order to determine the existence of a Bayes minimax rule, we define a new topology on the space of decision rules S. This new topology is intrinsic in the sense of Wald [Wald, 1950] and Teh Tjoe-Tie [Teh Tjoe-Tie, 1963]. However, our intrinsic topology differs in general from the topologies of Wald and Teh. Our new topology is defined as follows:

In the context of the given decision problem, let $r[\delta,\theta]$ denote the conditional risk function. Assume there is a decision rule δ_o with bounded risk i.e.,

$$\sup_{\theta\varepsilon R^n} r[\delta_o,\theta] \leq \beta < \infty .$$

Define $D \subset S$ by

$$D \overset{\Delta}{=} \{\delta\varepsilon S: \sup_{\theta\varepsilon R^n} r[\delta,\theta] \leq \beta\} .$$

Further, define $d: D\times D \to R$ by

$$d(\delta_1,\delta_2) \overset{\Delta}{=} \left|\sup_{\theta\varepsilon R^n} r[\delta_1,\theta] - \sup_{\theta\varepsilon R^n} r[\delta_2,\theta]\right| .$$

Observe that d is a metric on D under the usual equivalence class definition of points. We now define our intrinsic topology (I-topology) on D as the topology induced by the metric d.

We note that in general, our I-topology differs from the intrinsic topologies of Wald and Teh. In particular, Wald's intrinsic topology was induced by the metric

$$\sup_{\theta\varepsilon R^n} | r[\delta_1,\theta] - r[\delta_2,\theta]| ;$$

whereas Teh defined a base for his topology on S by means of the following subsets

$$V(\delta',\alpha) \overset{\Delta}{=} \{\delta\varepsilon S: \sup_{\theta\varepsilon R^n} [r[\delta',\theta] - r[\delta,\theta]] < \alpha\}$$

where $\delta'\varepsilon S$ and $\alpha > 0$.

We remark that:

(i) If $\omega(\delta) = \sup_{\theta\varepsilon R^n} r[\delta,\theta]$, then $\omega(\delta)$ is continuous in the I-topology on D.

(ii) Under the previous assumptions about L_1 and L_2 it can be shown that D is I-compact. Hence the $\inf_{\delta \varepsilon D} \sup_{\theta \varepsilon R^n} r[\delta,\theta]$ is achieved by some rule $\delta^* \varepsilon D$.

Our intrinsic topology was developed previously in [Mintz and Bock, 1975] to obtain saddle-point results for statistical games with noncompact strategy spaces and loss functions that exhibit a difference structure, i.e.,

$$L(\delta,\theta) = L_1(\delta-\theta) - L_2(\theta) .$$

Here, as in the present problem setting, L_1 and L_2 must satisfy certain continuity, convexity, and/or growth conditions. Since the results of this present paper make use of many of the ideas found in [Mintz and Bock, 1975], the reader will be referred to this earlier work to avoid duplication.

2. Main Results:

In this section we define the recurring terminology which appears in this paper, and list the underlying assumptions with respect to L_1, L_2, and F, on which the fundamental results of this paper rest. We then pose the statistical game and state our results as theorems 1-3.

2.1 Definitions:

$\mathcal{D}_1$. $\mathcal{F} \triangleq \{F(Z-\theta) : \theta \varepsilon R^n\}$; F, a given distribution function on R^n.

$\mathcal{D}_2$. F_θ; notation for $F(Z-\theta)$.

$\mathcal{D}_3$. $L(\delta,\theta)$; the generic loss function.

$L: R^n \times R^n \rightarrow R,$

$L(\delta,\theta) \triangleq L_1(\delta-\theta)L_2(\theta),$

$L_1: R^n \rightarrow R, \quad L_2: R^n \rightarrow R.$

$\mathcal{D}_4$. M; the set of all distribution functions on the Borel measurable space $(R^n,\mathcal{B}^n)$.

$\mathcal{D}_5$. $M(\rho) \triangleq \{\lambda \varepsilon M: \int ||\theta||^2 d\lambda \leq \rho\}$

$\mathcal{D}_6$. S; the set of all Borel maps taking $R^n \rightarrow R^n$.

$\mathcal{D}_7$. $\bar{\ell}_1 \triangleq \int L_1(x)dF(x),$

$\beta \triangleq \sup_\theta (\bar{\ell}_1 L_2(\theta))$

$\mathcal{D}_8$. $D \triangleq \{\delta \varepsilon S: \int L(\delta,\theta)dF_\theta \leq \beta, \text{ for all } \theta \varepsilon R^n\}.$

$\mathcal{D}_9$. $r[\delta,\theta] \triangleq \int L(\delta,\theta)dF_\theta$, $\delta\varepsilon D$, $\theta\varepsilon R^n$.

$\mathcal{D}_{10}$. $R[\delta,\lambda] = \int r[\delta,\theta]d\lambda$, $\delta\varepsilon D$, $\lambda\ \varepsilon \mathcal{M}$.

2.2 Assumptions:

A_1. L_1 is continuous.

A_2. L_1 is bounded below, i.e., $\inf\limits_x L_1(x) \triangleq \inf L_1 > -\infty$.

A_3. $\bar{\ell}_1 \triangleq \int L_1(x)dF(x) < \infty$.

A_4. For each $\delta\varepsilon D$,

$\int L_1(\delta(Z) - \theta)dF(Z-\theta)$

is continuous in θ, $\theta\varepsilon R^n$.

A_5. L_1 is convex.

A_6. L_2 is continuous.

A_7. L_2 is nonnegative.

A_8. $\beta \triangleq \sup\limits_\theta(\bar{\ell}_1\ L_2(\theta)) < \infty$

A_9. There exists ρ_o such that $\lambda \not\varepsilon \mathcal{M}(\rho_o)$ implies:

$$\bar{\ell}_1 \int L_2 d\lambda < \sup_{||\theta||^2 \leq \rho_o} \inf_x L_1(x)L_2(\theta).$$

2.3 Problem Statement:

Let Z denote a random vector, where $Z \sim F(Z-\theta)$. Using definitions $\mathcal{D}_1$-$\mathcal{D}_{10}$ and assumptions A_1-A_9 show that:

(i) The game admits a least favorable prior distribution λ^* for $\theta\varepsilon R^n$.

(ii) There is a minimax decision rule $\delta^*\ \varepsilon D$.

(iii) The minimax rule δ^* is D-Bayes with respect to λ^*.

We now state these objectives precisely as the following three theorems.

2.4 Theorems 1-3:

Consider the statistical decision problem determined by $L(\delta,\theta)$, $\delta\varepsilon D$, $\theta\varepsilon R^n$, and F, subject to assumptions A_1-A_9,

Theorem 1.

There exists a measure $\lambda^* \varepsilon M$ such that

$$\inf_{\delta\varepsilon D} R[\delta, \lambda^*] = \sup_{\lambda\varepsilon M}\, \inf_{\delta\varepsilon D} R[\delta,\lambda],$$

i.e., λ^* is least favorable.

Theorem 2.

There exists a decision rule $\delta^* \varepsilon D$ such that

$$\sup_{\theta\varepsilon R^n} r[\delta^*,\theta] = \inf_{\delta\varepsilon D}\, \sup_{\theta\varepsilon R^n} r[\delta,\theta].$$

i.e. δ^* is minimax.

Theorem 3.

Let λ^* and δ^* be defined by theorems 1 and 2. Then

$$R[\delta^*,\lambda^*] = \inf_{\delta\varepsilon D} R[\delta, \lambda^*],$$

i.e., δ^* is D-Bayes with respect to λ^*.

The proofs of these three theorems are contained in section 3.

Remarks:

(i) In our search for minimax rules, it suffices to restrict our attention to the (nonempty) set $D \subset S$, since

$$D \overset{\Delta}{=} \{\delta\varepsilon S\colon \sup_{\theta\varepsilon R^n} r[\delta,\theta] \leq \beta\}.$$

Hence, if $\delta' \notin D$, then there exists a rule $\delta\varepsilon D$, such that

$$\sup_{\theta\varepsilon R^n} r[\delta,\theta] < \sup_{\theta\varepsilon R^n} r[\delta',\theta].$$

Hence δ' cannot be minimax over S.

(ii) Sufficient conditions (with respect to L_1 and F) to guarantee A_4 are discussed in [Mintz and Bock, 1975].

3. Proofs of Theorems 1-3:

Our proofs of theorems 1-3 are based on a sequence of observations and lemmas which we develop in this section. The proofs of several of these preliminary lemmas are in turn based on some results in Billingsley, 1968; Breiman, 1968; Ferguson, 1967; and Sion, 1958. These results (or immediate extensions) are stated in the appendix to [Mintz and Bock, 1975].

3.1 Proof of Theorem 1:

Observation 1. $M(\rho)$ is a convex subset of M.

Lemma 1. If $\rho < \infty$, then $M(\rho)$ is compact in the weak topology on M.

Proof: See [Mintz and Bock, 1975], lemma 1.

Observation 2. The set of decision rules D is convex.

Observation 3. For each $\delta \varepsilon D$, $R[\delta,\lambda]$ is concave in λ, $\lambda \varepsilon M$.

Observation 4. For each $\lambda \varepsilon M$, $R[\delta,\lambda]$ is convex in δ, $\delta \varepsilon D$.

Lemma 2. For each $\delta \varepsilon D$, $r[\delta,\theta]$ is continuous in θ, $\theta \varepsilon R^n$.

Proof: The θ continuity of

$$r[\delta,\theta] = \int L_1(\delta(Z) - \theta) dF(Z-\theta) L_2(\theta)$$

for each $\delta \varepsilon D$ follows directly from assumptions A_4 and A_6.

Lemma 3. For each $\delta \varepsilon D$, $R[\delta,\lambda]$ is upper semicontinuous on M in the topology of weak convergence (w-usc).

Proof: The proof of lemma 3 is nearly identical to that of lemma 3 in [Mintz and Bock, 1975]. We merely note that for each $\delta \varepsilon D$, $-r[\delta,\theta]$ is bounded below and continuous in θ as a consequence of assumptions A_4, A_6, and A_8, and definition D_8.

Lemma 4. For each $\lambda \varepsilon M$, $R[\delta,\lambda]$ is lower semicontinuous on D in the topology of pointwise convergence (pw-lsc).

Proof: The proof is developed in several steps.

(i) To show $R[\delta,\lambda]$ is pw-lsc in δ, it is sufficient to show that

$$\iint L_1(\delta-\theta) dF_\theta L_2(\theta) d\lambda$$

is pw-lsc in δ for fixed λ, since L_2 does not depend on δ.

(ii) Fix $\lambda \varepsilon M$, and assume $\delta_n \overset{pw}{\rightarrow} \delta$. We wish to show:

$$\underline{\lim}_n \iint L_1(\delta_n-\theta) dF_\theta L_2(\theta) d\lambda \geq \iint L_1(\delta-\theta) dF_\theta L_2(\theta) d\lambda.$$

Since L_1 is bounded below (assumption A_2), and $L_2 \geq 0$ (assumption A_7) then by Fatou's lemma it is sufficient to show that for all $\theta \varepsilon R^n$:

$$\underline{\lim}_n \int L_1(\delta_n-\theta) dF_\theta \geq \int L_1(\delta-\theta) dF_\theta.$$

This latter inequality follows by a second application of Fatou's lemma, the continuity of L_1, and the assumption of pointwise convergence, namely, for all $\theta \varepsilon R^n$,

$$\underline{\lim}_n \int L_1(\delta_n-\theta) dF_\theta \geq \int \underline{\lim}_n L_1(\delta_n-\theta) dF_\theta$$

$$= \int L_1(\delta-\theta) dF_\theta .$$

Hence for each $\lambda \varepsilon M$, $R[\delta,\lambda]$ is pw-lsc on D.

Lemma 5. If $0 < \rho < \infty$, then:

$$\inf_{\delta \varepsilon D} \sup_{\lambda \varepsilon M(\rho)} R[\delta,\lambda] = \sup_{\lambda \varepsilon M(\rho)} \inf_{\delta \varepsilon D} R[\delta,\lambda].$$

Proof: Using lemmas 1, 3, and 4; and observations 1-4: it follows that

(i) D is convex, and $M(\rho)$ is convex and weakly compact.

(ii) For fixed $\delta \varepsilon D$, $R[\delta,\lambda]$ is concave and weakly upper semicontinuous on $M(\rho)$.

(iii) For fixed $\lambda \varepsilon M(\rho)$, $R[\delta,\lambda]$ is convex and lower semicontinuous in the pointwise topology on D.

Hence by Sion's theorem [Sion, 1958, corollary 3.3, p. 174], the proof is complete.

Lemma 6. If $0 < \rho < \infty$, then:

$$\inf_{\delta \varepsilon D} \max_{\lambda \varepsilon M(\rho)} R[\delta,\lambda] = \max_{\lambda \varepsilon M(\rho)} \inf_{\delta \varepsilon D} R[\delta,\lambda] .$$

Proof: See [Mintz and Bock, 1975], lemma 6.

Lemma 7. Let ρ_o satisfy assumption A_9. If $\rho > \rho_o$, then all maximizing solutions $\lambda^*(\rho)$ for the saddle-point problem:

$$\inf_{\delta \varepsilon D} \max_{\lambda \varepsilon M(\rho)} R[\delta,\lambda] = \max_{\lambda \varepsilon M(\rho)} \inf_{\delta \varepsilon D} R[\delta,\lambda]$$

must lie in $M(\rho_o)$.

Proof: [By contradiction].

We develop the proof in several steps:

(i) Let $V(\rho) = \max_{\lambda \varepsilon M(\rho)} \inf_{\delta \varepsilon D} R[\delta,\lambda]$,

and observe that

$$\sup_{||\theta||^2 \leq \rho_o} \inf_x L_1(x)L_2(\theta) \leq V(\rho).$$

(ii) If $\delta_o \triangleq \delta_o(Z) \equiv Z$, then $\delta_o \varepsilon D$ and $R[\delta_0,\lambda] = \bar{\ell}_1 \int L_2 d\lambda$.

(iii) We obtain a contradiction as follows: Suppose the contrary, i.e., $\lambda^*(\rho) \ \varepsilon M(\rho) - M(\rho_o)$. Then by assumption A_9, we have

$$R[\delta_o,\lambda^*] = \bar{\ell}_1 \int L_2 d\lambda < V(\rho),$$

which contradicts the hypothesis that

$$V(\rho) = \max_{\lambda \varepsilon M(\rho)} \inf_{\delta \varepsilon D} R[\delta,\lambda].$$

As a direct consequence of lemma 7 we have:

Observation 5. $\inf_{\delta \varepsilon D} \max_{\lambda \varepsilon M} R[\delta,\lambda] = \max_{\lambda \varepsilon M} \inf_{\delta \varepsilon D} R[\delta,\lambda]$, i.e., referring to the saddle-point result in lemma 6, $M(\rho)$ can be replaced by M when ρ is sufficiently large.

Theorem 1.

Consider the statistical decision problem determined by $L(\delta,\theta)$, $\delta \varepsilon D$, $\theta \varepsilon R^n$, and F, subject to assumptions A_1-A_9. Then there exists a least favorable prior distribution $\lambda^* \varepsilon M$.

Proof: By definition, a prior distribution λ^* is least favorable over M if

$$\inf_{\delta \varepsilon D} \int r[\delta,\theta] d\lambda^* = \sup_{\lambda \varepsilon M} \inf_{\delta \varepsilon D} \int r[\delta,\theta] d\lambda.$$

The existence of λ^* is established by observation 5.

3.2 Proof of Theorem 2:

Observation 6. Let $\omega(\delta) \overset{\Delta}{=} \sup r[\delta,\theta]$, $\delta \varepsilon D$. Now with reference to the definition of D, $(\mathcal{D}_8)$, we note that

$$D = \{\delta \varepsilon S\colon \omega(\delta) \leq \beta\}.$$

Observation 7. The function $d\colon D x D \to R$,

$$d(\delta_1,\delta_2) = |\omega(\delta_1) - \omega(\delta_2)|,$$

is a pseudometric on D. Further, $d(\cdot,\cdot)$ becomes a metric if we make the usual equivalence class identifications.

Definition: Let I denote the intrinsic topology induced on D by the metric d. In what follows we will show that:

(i) (D,d) is a complete metric space.

(ii) D is compact in the I-topology.

(iii) $\omega(\delta)$ is continuous in the I-topology on D.

Results (i)-(iii) will insure the existence of a minimax decision rule.

Lemma 8. (D,d) is a complete metric space.

Proof: See [Mintz and Bock, 1975], lemma 8, where β now replaces c.

Observation 8. For all $\delta \varepsilon D$,

$$\inf L_1 \leq \omega(\delta) \leq \beta.$$

Lemma 9. D is compact in the I-topology.

Proof: See [Mintz and Bock, 1975], lemma 9.

Observation 9. $\omega(\delta)$ is I-continuous on D.

Theorem 2.

There exists a decision rule $\delta^* \varepsilon D$ such that:

$$\sup_{\theta \varepsilon R^n} r[\delta^*,\theta] = \inf_{\delta \varepsilon D} \sup_{\theta \varepsilon R^n} r[\delta,\theta].$$

Proof: By lemma 9 and observation 9, D is I-compact and ω is I-continuous on D. Hence it follows that ω achieves a minimum value on D for some element $\delta^* \varepsilon D$, i.e.,

$$\omega(\delta^*) = \inf_{\delta \varepsilon D} \omega(\delta).$$

Therefore δ^* is D minimax.

3.3 Proof of Theorem 3:

Theorem 3.

Referring to the given decision problem and theorems 1 and 2, let δ^* denote a minimax rule over D and λ^* a least favorable prior distribution over M. Then,

$$R[\delta^*,\lambda^*] = \inf_{\delta \varepsilon D} R[\delta,\lambda^*].$$

Proof: See [Mintz and Bock, 1975], theorem 3.

4. Applications To LEG Stochastic Differential Games:

4.1 An Estimation Problem:

Let $L^+(\delta,u)$ denote the following loss function:

$$L^+(\delta,u) = \exp(||x(T) - \delta(\underline{Z})||_C^2)\exp(-\int_0^T u'Dudt),$$

where x(T) denotes the terminal state of a given linear system with associated observations $\underline{Z}$. Our goal here is to find a minimax estimate for x(T) given $\underline{Z}$, under the exponentiated generalized quadratic loss function $L^+(\delta,u)$.

This minimax state estimation problem can be viewed as an example of the function space generalization of the class of statistical games posed in section 2. The basic ideas in sections 2 and 3 can be extended to encompass this function space setting. This extension to a function space setting including a focus on the linear exponential gaussian (LEG) examples is the subject of a forthcoming paper [Bock, Padilla, and Mintz, 1977].

References

See [Mintz and Bock, 1975] for reference information for Billingsley, 1968; Breiman, 1968; Ferguson, 1967; Sion, 1958; Strawderman, 1971; Teh, 1963; and Wald, 1950.

1. Bock, M. E., C. Padilla, and M. Mintz, "On LEG Stochastic Differential Games," forthcoming, 1977.
2. Mintz, M. and M. E. Bock, "On the Existence of Saddle-Point Solutions for Statistical Games with Noncompact Strategy Spaces," Proceedings of the 1975 IEEE Conference On Decision and Control, pp. 615-621.
3. Stelte, D. and M. Mintz, "On Minimax Parameter Estimation Under Exponentiated Generalized Quadratic Loss," Proceedings of the Eleventh Allerton Conference On Circuit and System Theory, 1973, pp. 215-225.

OPTIMAL CONTROL OF TWO-LEVEL DYNAMIC SYSTEMS

M. SIMAAN
Department of Electrical Engineering
University of Pittsburgh
Pittsburgh, Pennsylvania 15261

Abstract

This paper presents an approach, based on the Stackelberg solution of game theory, for the optimization of dynamic systems upon which control is exerted at two different levels. At one level, there is a central controller whose decisions are transmitted to a lower level at which there are N subsystems each equipped with its own local controller. Each local controller has an objective function which assesses the individual performance of its corresponding subsystem, and the central controller has an objective function which is, in general, different from that of any of the subsystems and which assesses the global performance of the overall system. The Stackelberg concept is used for the performance optimization of such systems.

1. Introduction

Recently, since the appearance of Starr and Ho's results [1-3] on nonzero-sum differential games (or generalized control theory), there has been considerable emphasis in the control and optimization literature on the study of systems that are controlled by more than one controller each having a different objective function. Several solution concepts for such systems have since been proposed and studied. Some of these concepts, such as Nash and Minimax [1-3] describe a noncooperative situation among the various controllers. Others, such as the Pareto solution [4] describe a cooperative situation among the controllers. Another concept which has been studied by Chen and Cruz [5] and Simaan and Cruz [6-9] is suitable for situations where there are two players (or controllers), but their roles are such that one is a leader who announces his strategy first and the other is a follower who announces his strategy second.

At the same time, motivated by problems involving large scale power networks, industrial chemical plants and metallurgical processes as well as economic and social systems, recently there has been considerable emphasis in the control and optimization literature on the study of Large Scale Systems. The literature on this subject is vast and we only refer to a recent survey by Sandell, Varaiya and Athans [10] which summarizes recent trends in this field.

Clearly, generalized control theory and large scale system theory have a lot in common. Large scale systems are seldom controlled by one controller and seldom all controllers have the same performance indices. The decentralized structure is basic to the study of Large Scale Systems.

In this paper, the special case of two-level decentralized large scale systems is studied within the framework of generalized control theory. At the high level, we assume that there is a central coordinator (or controller) whose decisions are transmitted to the lower level at which there are N subsystems each with its own local independent controller. We assume that each local controller has an objective function which assesses the performance of its corresponding subsystem, and that the central coordinator has an objective function which is, in general different from that of any of the subsystems and which assesses the (global) performance of the overall system. The Stackelberg concept of generalized control theory [6] is considered for the performance optimization of such systems.

2. Formulation of Two-Level Optimal Control Problems

For the sake of simplicity, and whenever analytic solutions are to be derived, we shall assume that the system under consideration is linear and that all objective functions are quadratic. The concepts developed, however, are general and apply to nonlinear as well as to static two-level optimization problems.

Consider a large dynamic system (see figure 1) which is composed of N subsystems, each described by a state equation of the form:

$$\dot{x}_i = A_i x_i + B_i u_i + C_i v \qquad ,i = 1,\ldots,N \tag{1}$$

where x_i is the state vector of dimension n_i of subsystem i; and u_i is the local control vector of dimension m_i of subsystem i. The control vector v of dimension m_o, which is common to all subsystems is the control of the central coordinator. The matrices A_i, B_i and C_i are of dimensions $n_i x n_i$, $n_i x m_i$ and $n_i x m_o$ respectively. As is evident form (1), we have chosen to treat the subsystems as connected only through the control v of the coordinator. No interconnection is assumed to exist through their state variables. The reason for this is that if the subsystems were assumed to be interconnected through their state variables, their operation will then depend on controls of other local controllers and the result will be a more complicated differential game situation among the subsystems. This is not the main theme of this paper, and research on the differential game aspect of two level Large Scale Systems will be the subject of a further study. Nevertheless, if such coupling exists and is weak, then by using an ε-coupling technique [11], a first order approximation of the solution of the weakly coupled differential game can then be obtained by solving the uncoupled sybsystem case, discussed in this paper, as a first step of the approximation.

Furthermore, intuitively our starting point is fundamental to the interconnectibility of large scale systems. For instance, if the coordinator chooses a feedback control of the form $v = \sum_{j=1}^{N} L_j x_j$ where the matrices L_i, $i = 1,\ldots,N$ are chosen either arbitrarily or based on some optimality criterion, then the result will be an interconnected large scale system:

$$\dot{x}_i = A_{ii} x_i + \sum_{\substack{j=1 \\ j \neq i}}^{N} A_{ij} x_j + B_i u_i \tag{2}$$

where $A_{ii} = A_i + C_i L_i$ and $A_{ij} = C_i L_j$, of the type considered in the large scale systems literature [10]. Thus, interconnectibility among subsystems may be the result of a feedback control by the central coordinator. However, it is interesting to note that the system may then be re-decoupled by the use of the local controllers.

We shall assume that at the local level, each subsystem has a performance index:

$$J_i(u_i,v) = 1/2\, x_i'(t_f) K_{if} x_i(t_f) + 1/2 \int_{t_o}^{t_f} (x_i' Q_i x_i + u_i' R_i u_i + v' P_i v)dt \tag{3}$$

where K_{if}, Q_i, R_i and P_i are symmetric matrices of dim. $n_i x n_i, n_i x n_i, m_i x m_i$ and $m_o x m_o$ respectively. The term $v'P_iv$ may be set equal to zero in some cases where the subsystems are controlled solely at the local level. The central controller, who is concerned with the performance of the overall system has a performance index

$$J_o(u_1,\dots,u_N,v) = 1/2\ x'(t_f)K_{of}x(t_f) + 1/2 \int_{t_o}^{t_f} (x'Q_ox + u'R_ou + v'P_ov)dt \quad (4)$$

where $x' = (x_1', x_2',\dots,x_N')$ is the state vector of the overall system and $u' = (u_1',u_2',\dots,u_N')$ is the vector of all local controls. The symmetric matrices K_{of}, Q_o, R_o and P_o are $n \times n$, $n \times n$, $m \times m$ and $m_o \times m_o$ respectively, where $n = \sum_{i=1}^{N} n_i$ and $m = \sum_{i=1}^{N} m_i$. All matrices in (1)-(4) except for K_{if} and K_{of} are assumed to be time varying.

We should note at this stage that in general the central coordinator cannot take over the system by cancelling the controls of the local controllers. Only under two very special cases this may be possible:

i) when the controls of the local controllers are nullified and the coordinator takes over (as discussed in case b below)

ii) when the system dynamics is such that there exists matrices D_i, $i = 1,\dots,N$ which satisfy $\forall\ t \in [t_o,t_f)$ the equations $B_i u_i + C_i v = D_i v$, $i = 1,\dots N$.

The fact that the coordinator cannot in general take over the system is important because otherwise the decentralized structure of the system will be destroyed. Furthermore we note that even though the coordinator is assumed to have knowledge of the objective functions of the subsystems; each subsystem does not necessarily have to know the objective function of the coordinator, nor that of other subsystems.

The performance index of the coordinator can, by a proper choice of the matrices K_{of}, Q_o, R_o and P_o be made equal to a convex sum of the performance indices of the subsystems. This is achieved if:

$$K_{of} = \text{block diag. } \{\alpha_1 K_{1f}, \alpha_2 K_{2f},\dots,\alpha_N K_{Nf}\}$$

$$Q_o = \text{block diag. } \{\alpha_1 Q_1, \alpha_2 Q_2,\dots, \alpha_N Q_N\}$$

$$R_o = \text{block diag. } \{\alpha_1 R_1, \alpha_2 R_2,\dots, \alpha_N R_N\}$$

and $P_o = \sum_{i=1}^{N} \alpha_i P_i$, where the vector $\alpha = (\alpha_1,\dots,\alpha_N)'$ is $\in M$;

$$M = \{\alpha \in R^N : \sum_{i=1}^{N} \alpha_i = 1 \text{ and } \alpha_i \geq 0\} \quad (6)$$

In this paper, the above format (5) - (6) is not necessarily imposed, however as we shall describe later, it may have some interesting physical significance, and will result in considerable computational savings.

Since the system under consideration has controls exerted at two levels, the following three cases will be considered:

Case a: Control Exerted Only by Local Controllers. If the coordinator's control is nullified ($v \equiv o$), the subsystems become uncoupled; and each one solves an ordinary linear quadratic problem. If $R_i > 0$ and

$Q_i \geq 0$, the optimal local control of the i^{th} subsystem is $u_i^* = -R_i^{-1} B_i' K_i(t)\, x_i(t)$ where $K_i(t)$ is the positive definite solution of the i^{th} subsystem's Riccati differential equation:

$$\dot{K}_i = -A_i'K_i - K_iA_i - Q_i + K_i B_iR_i^{-1}B_i'K_i, \quad K_i(t_f) = K_{if} \quad . \tag{7}$$

The i^{th} subsystem's trajectory satisfies $\dot{x}_i = (A_i - B_iR_i^{-1}B_i'K_i)x_i$, $x_i(t_o) = x_{io}$, and its optimal performance is measured as: $J_i^* = 1/2\ x_{io}'\ K_i(t_o)x_{io}$.

The performance of the overall system is computed as $J_o^* = 1/2\ x_o'\ F(t_o)\ x_o$ where the n x n matrix F(t) is the solution of the linear matrix differential equation:

$$\dot{F} = -\hat{A}'F - F\hat{A} - \hat{Q} \ , \quad F(t_f) = K_{of} \tag{8}$$

where

$$\hat{A} = \text{block diag. } \{A_1 - B_1R_1^{-1}B_1'K_1,\ A_2 - B_2R_2^{-1}B_2'K_2, \ldots, A_N - B_NR_N^{-1}B_N'K_N\},$$

$$\hat{Q} = Q_o + K_o'R_oK_o, \text{ with}$$

$$K_o = \text{block diag. } \{R_1^{-1}B_1'K_1,\ R_2^{-1}B_2^1\ K_2, \ldots, R_N^{-1}B_N'K_N\} \tag{9}$$

Obviously even though each subsystem is operating optimally, the performance of the overall system, from the coordinator's point of view may not be optimal.

We note also that for the time invariant, infinite horizon case, the stability of the overall system is a result of the stability of all subsystem individually. Furthermore, because of decoupling, failures at a subsystem will not affect other subsystems.

Case b: Control Exerted Only by the Central Controller. If the controls of all local controllers are nullified (i.e. $u_i \equiv 0$, i = 1, ... N), then all subsystems are controlled by the central coordinator:

$$\dot{x}_i = A_ix_i + C_iv \tag{10}$$

whose performance index (4) is

$$J_o = 1/2\ x'(t_f)K_{of}\ x(t_f) + 1/2 \int_{t_o}^{t_f} (x'Q_ox + v'P_ov)dt \tag{11}$$

Obviously the coordinator cannot minimize all J_i's simultaneously. If J_o is set equal to J_{i_o} for some subsystem i_o, then only one subsystem can be operated optimally, and all other subsystems will operate in a non-optimal manner. If, on the other hand, J_o is set equal to a convex combination of all other J_i's as in (5) - (6):

$$J_o = \tilde{J}_o(v,\alpha) = \sum_{i=1}^{N} \alpha_iJ_i \quad \text{with } \alpha \in M \tag{12}$$

Then the optimal solution will be α-dependent. The choice of the vector $\alpha = (\alpha_1, \ldots, \alpha_N)'$ can be either arbitrary in M or based on some rationale. The Minimax Pareto approach considered by Medanic et al. [12] for multiple target problems seems to be very effective in this case. If v* and α* are chosen such that

$$\tilde{J}_o(v^*,\alpha^*) = \max_{\alpha\epsilon M} \min_{v} \tilde{J}_o(v,\alpha) = \min_{v} \max_{\alpha\epsilon M} \tilde{J}_o(v,\alpha) \tag{13}$$

where a Saddle point has been shown to exist [12], then (v^*,α^*) has the property that v^* is a noninferior control and that

$$\tilde{J}_o(v^*,\sigma^*) = \min_{v} \max_{i} J_i \qquad i = 1, \ldots, N \tag{14}$$

where all $u_i \equiv 0$ in J_i. The significance of (14) is that the subsystem whose performance is worst, gets helped most by the coordinator.

The more general case, where J_o is not a convex combination of the J_i's, can also be considered. The result, however, will not in general be optimal for any of the subsystems. The optimal control then requires measurements of all states and (assuming that $P_o > 0$ and $Q_o \geq 0$) is given by: $v = -P_o^{-1}C'Ex$ where the n x n matrix E is the solution of the overall system's Riccati equation:

$$\dot{E} = -A'E - EA - Q_o + EC'P_o^{-1}CE \quad , \quad E(t_f) = K_{of} \tag{15}$$

with

$$A = \text{block diag.}\ \{A_1, A_2, \ldots, A_N\} \tag{16}$$

and

$$C = [C_1' \mid C_2' \mid \cdots \mid C_N']' \tag{17}$$

We note that even if Q_o and P_o were block diagonal matrices; equation (15) will not decouple into N Riccati Equations of order n_i.

Furthermore, we mention that in the infinite horizon case as a result of (15), the stability of the overall system becomes very much a function of the operation of all subsystems. If a failure occurs at one subsystem (for instance due to a loss of control from the coordinator) then the overall system may become unstable; (thus a decrease in reliability).

Case c: Control Exerted Simultaneously by the Central and Local Controllers. In this case, control is assumed to be provided at both levels, the coordinator level and the local controllers level (see figure 1). The system's equation and performance indices are as in (1), (3) and (4). Since each controller has his own performance index, our approach to formulating a solution for this case will be based on the concept of Stackelberg equilibrium solutions of game theory [13]. This concept which was first developed for static duopoly markets, has recently been extended to two player differential games [5-7]. A study of this solution for the case of two groups of N players has been presented in [8] and more recently a development of Stochastic Stackelberg solutions has been given in [14].

The Stackelberg solution seems to be the natural approach for solving the two-level system (1), (3) and (4), under consideration in this paper. It is reasonable to assume that, the central coordinator, before deciding on his control, will determine how the subsystems will be optimized at the local level once his decisions are transmitted. He then computes his control in such a way that, once transmitted to the subsystems, the corresponding optimal local controllers will have the property that the overall performance of the system will also be optimized. A definition of this solution such that it is directly applicable to the two-level structure under consideration is given below. Let U_i, $i = 1, \ldots, N$ be the admissible control sets for the local controllers and let V be the admissible control set for the coordinator. Let $J_i(u_i,v)$, $i = 1,\ldots,N$ be the objective functions of the local controllers, J_i: $U_i X V \to R$ and let $J_o(u_1,\ldots,u_N,v)$ be the objective function of the coordinator, J_o: $\prod_{i=1}^{N} U_i \, XV \to R$. Whether the system is static or dynamic, we will assume that there exists mappings T_i, $i=1,\dots,N$ such that T_i: $V \to U_i$ such that for any $v \in V$ and for $i = 1,\dots,N$,

$$J_i(T_i v, v) \leq J_i(u_i, v) \quad \forall \quad u_i \in U_i \tag{18}$$

These mappings determine how each controller will react once the coordinator has determined his control. We shall assume that these mappings are well defined and unique.

Definition: If there exists a control $v_s \in V$ such that

$$J_o(T_1 v_s, T_2 v_s, \dots, T_N v_s, v_s) \leq J_o\,(T_1 v, T_2 v, \dots, T_N v, v) \tag{19}$$

for all $v \in V$, then v_s is called the Stackelberg Control of the coordinator and the control set $(u_{1s}, u_{2s}, \dots, u_{Ns}, v_s)$ where $u_{is} = T_i v_s$ is called the Stackelberg solution of the system.

More specifically, in this solution the coordinator anticipates how the subsystems will react for a given decision made by him and he then optimizes the performance of the overall system accordingly. The subsystems will then react by optimizing their local performances independently but simultaneously.

As in case (b), if $J_o = J_{i_o}$ for some subsystem i_o, then the coordinator's objective is to help subsystem i_o to achieve best performance under both control levels. It can be shown in this case that the Stackelberg solution leads to a minimum of J_{i_o} with respect to all $u_{i_o} \in U_{i_o}$ and all $v \in V$. Other subsystems, however, will only be optimal against the control of the coordinator.

On the other hand, if the objective function of the coordinator is a convex combination of the local objective functions, that is if

$$J_o(u_1, \dots, u_N, v) = \check{J}_o(\alpha, u_1, \dots, u_N, v) \triangleq \sum_{i=1}^{N} \alpha_i J_i(u_i, v) \tag{20}$$

where the vector $\alpha \in M$, then the corresponding Stackelberg solution $(u_{1s\alpha}, \dots, u_{Ns\alpha}, v_{s\alpha})$ will also be a member of the noninferior set of $\{J_1(u_1, v), \dots, J_N(u_N, v)\}$. To show this, from (19) we have

$$\sum_{i=1}^{N} \alpha_i J_i(T_i v_{s\alpha}, v_{s\alpha}) \leq \sum_{i=1}^{N} \alpha_i J_i(T_i v, v) \quad \forall \; v \in V. \tag{21}$$

And from (18) it follows that

$$\sum_{i=1}^{N} \alpha_i J_i(T_i v, v) \leq \sum_{i=1}^{N} \alpha_i J_i(u_i, v) \tag{22}$$

for all $u_i \in U_i$, $i=1,\dots,N$. Combining (21) and (22), the Stackelberg solution is clearly a noninferior solution.

Now suppose that $\alpha^* \in M$ and a $v_{s\alpha^*} \in V$ are chosen such that:

$$\check{J}_o(\alpha^*, T_1 v_{s\alpha^*}, \dots, T_N v_{s\alpha^*}, v_{s\alpha^*}) = \max_{\alpha \in M} \min_{v \in V} \check{J}_o(\alpha, T_1 v, \dots T_N v, v) \tag{23}$$

then (23) has a saddle point [12], and we have:

$$\check{J}_o(\alpha^*, T_1 v_{s\alpha^*}, \dots, T_N v_{s\alpha^*}, v_{s\alpha^*}) = \min_{v \varepsilon V} \max_{\alpha \varepsilon M} \check{J}_o(\alpha, T_1 v, \dots, T_N v, v) \quad . \tag{24}$$

Now by using the same theorem [12] used to prove (16) (with a slight, simple modification) we have

$$\max_{\alpha \varepsilon M} \check{J}_o(\alpha, T_1 v, \dots, T_N v, v) = \max_{i} J_i(T_i v, v) \tag{25}$$

Therefore, this selection of α^* can be visualized as locating a Stackelberg point in the noninferior set, where the coordinator has chosen $v_{s\alpha^*}$ such that the maximum of $\{J_1(T_1 v_{s\alpha^*}, v_{s\alpha^*}), \dots, J_N(T_N v_{s\alpha^*}, v_{s\alpha^*})\}$ has been minimized. Therefore, this Stackelberg solution has the interesting property that the coordinator determines his control in such a way that, when all subsystems react optimally, the subsystem that gets hurt most, gets helped most by the coordinator. In other words, the coordinator helps most the subsystem that has the worst performance and needs most help. Of course, optimality is eventually achieved at the local level by proper choice of the optimal local controls.

In the general case, when the coordinator's objective function is not related to the objective functions of the subsystems the coordinator must solve the two step optimal control problem (18) - (19) in order to determine v_s.

For a dynamic system, there are several forms according to which v_s can be determined. More specifically v_s can be either in open-loop or closed-loop form. Because of the nature of the information patterns, as is well known in the theory of multicontroller multiobjective systems, these two forms lead in general to different solutions. In this paper, we shall illustrate how the open-loop solution is obtained by deriving necessary conditions for Stackelberg optimality. The open-loop solution is important in cases where the coordinator does not have measurements of the state vectors $x_i(t)$ of all subsystems, but knows their initial starting conditions. This stems from practical considerations in situations where the subsystems are physically located such that the transfer of state information to the central controller is either impossible or impractical.

If $v_s(t)$ is transmitted to the subsystems, then the vector $u = (u_1', u_2', \dots u_N')'$ of all optimal local controllers which minimizes (3) for each i can be computed as follows:

$$u = -R^{-1} B' \lambda \tag{26}$$

where λ satisfies

$$\dot{\lambda} = -Qx - A'\lambda \quad , \quad \lambda(t_f) = K_f \, x(t_f) \tag{27}$$

and where

R = block diag. $\{R_1, R_2, \dots R_N\}$

B = block diag. $\{B_1, B_2, \dots B_N\}$

Q = block diag. $\{Q_1, Q_2, \dots Q_N\}$

A = block diag. $\{A_1, A_2, \dots A_N\}$

K_f= block diag. $\{K_{1f}, K_{2f}, \dots K_{Nf}\}$

If we now append (26) - (27) as constraints in minimizing (4) with respect to v, the following conditions are obtained:

$$\dot{p} = -Q_o x - A'p + Q\gamma \quad , \quad p(t_f) = K_{of}\, x(t_f) - K_f \gamma(t_f) \tag{28}$$

$$\dot{\gamma} = -BR^{-1}R_o R^{-1}B'\lambda + BR^{-1}B'p + A\gamma \quad , \quad \gamma(0) = 0 \tag{29}$$

and $v_s = -P_o^{-1}C'p$ (30)

It is now possible to solve the TPBVP (26) - (30) for v_s and u_s. The dependence of these controls on x_o, however, is explicitly apparent if we introduce the following transformations:

Let

$\lambda = K\, x \;,\; p = M\, x$ and $\gamma = Lx$, (31)

Then by substituting in (26) - (30), and omitting algebraic manipulations, the following system is obtained:

$$\dot{K} = -A'K - KA - Q + KBR^{-1}B'K + KCP_o^{-1}C'M \tag{32}$$

$$\dot{M} = -A'M - MA - Q_o + QL + MBR^{-1}B'K + MCP_o^{-1}C'M \tag{33}$$

$$\dot{L} = AL - LA + BR^{-1}B'M + LBR^{-1}B'K + LCP_o^{-1}C'M - BR^{-1}R_oR^{-1}B'K \tag{34}$$

with boundary conditions $K(t_f) = K_f$, $M(t_f) = K_{of} - K_f\, L(t_f)$ and $L(t_o) = 0$. The Stackelberg controls are:

$$\begin{aligned} u_s &= -R^{-1}B'K\, \phi(t,t_o)x_o \\ v_s &= -P_o^{-1}C'M\phi(t,t_o)x_o \end{aligned} \tag{35}$$

where ϕ is the state transition matrix of the overall system:

$$\dot{x} = (A - BR^{-1}B'K - CP_o^{-1}C'M)x \quad . \tag{36}$$

We note that all controls in (35) for the coordinator and local controllers are coupled through the initial conditions. This is the only coupling among the subsystems and it is due to the fact that since $v_s(t)$ is a function of all initial states, and since each $u_i(t)$ depends on $v_s(t)$ (4, 26, 27), then each $u_{is}(t)$ depends on all initial conditions. This dependence, however, is implicit since $v_s(t)$ is transmitted to the subsystems as a function of time once $x(t_o)$ is known by the coordinator.

The above Stackelberg approach is now illustrated by the following example:
Example: Consider a system with one coordinator v and two local controllers u_1 and u_2. Let the dynamics of the two subsystems, and the objective functions of their local controllers be of the form:

$$\dot{x}_1 = -u_1 + v,\; x_1(0) = x_{1o} \;,\; J_1 = -\frac{1}{2}\, x_{1f}^2 + \frac{1}{2c_1}\int_0^1 u_1^2\, dt$$

$$\dot{x}_2 = -u_2 + v\ ,\ x_2(0) = x_{2o},\ J_2 = -\frac{1}{2}\ x_{2f}^2 + \frac{1}{2c_2}\int_0^1 u_2^2\, dt$$

and let the objective function of the coordinator[†] be:

$$J_o = \frac{1}{2}\ (x_{1f} - x_{2f})^2 + \frac{1}{2c_o}\int_0^1 v^2\, dt$$

Assuming $0 < c_1, c_2 < 1$, and $c_o > 0$ and applying conditions (26) - (36) the open-loop Stackelberg solution for this two-level optimal control problem is:

$$v_s = \frac{\beta c_o}{1+\beta^2 c_o}\ (\frac{x_{2o}}{1-c_2} - \frac{x_{1o}}{1-c_1}),\ \text{where}\ \beta = \frac{c_1 - c_2}{(1-c_1)\ (1-c_2)}$$

$$u_{1s} = \frac{c_1}{c_{1}-1}\ [\frac{c_o\beta}{c_o\beta^2+1}\ \frac{x_{2o}}{1-c_2} + (1 - \frac{c_o\beta}{(c_o\beta^2+1)\ (1-c_1)})\ x_{1o}],\ \text{and}$$

$$u_{2s} = \frac{c_2}{c_{2}-1}\ [(1 + \frac{c_o\beta}{(c_o\beta^2+1)\ (1-c_2)})\ x_{2o} - \frac{c_o\beta}{c_o\beta^2+1}\ \frac{x_{1o}}{1-c_1}]$$

Note the dependence of v_s, u_{1s} and u_{2s} on x_{1o} and x_{2o}.

3. Conclusion

In this paper, the Stackelberg approach of game theory has been applied to the optimal control of two-level dynamic systems. These systems consist of a coordinator at the high level who transmitts decisions to N subsystems at the low level each equipped with its own local controller. The objective function of the coordinator assesses the global performance of the system while the objective function of each local controller assesses the performance of its corresponding subsystem. Various possibilities of solution have been considered and an example illustrating the Stackelberg approach has been given.

4. References

[1] Starr, A.W. and Ho, Y.C., "Nonzero-Sum Differential Games," Journal of Opti. Th. and App. (J.O.T.A.), Vol. 3, No. 3, 1969, pp. 184-206.

[2] Starr, A.W. and Ho, Y.C., "Further Properties of Nonzero-Sum Differential Games," J.O.T.A., Vol. 3, No. 4, 1969, pp. 207-219.

[3] Ho, Y.C., "Differential Games, Dynamic Optimization, and Generalized Control Theory," J.O.T.A.,Vol. 6, No. 3, 1970, pp. 179-209.

[4] Schmitendorf, W.E., "Cooperative Games and Vector-Valued Criteria Problems," IEEE Trans. on Aut. Cont. Vol. AC-18, No. 2, April 1973, pp. 139-144.

[†] Note that this problem can be interpreted as a one pursuer, two-evader version of the velocity controlled pursuit-evasion problem studied in [1] and [6]. The evaders u_1 and u_2 want to maximize their distances from the pursuer, while the pursuer wants to minimize the gap between the two evaders (in order to hit them simultaneously with one shot).

[5] Chen, C.I. and Cruz, J.B., Jr., "Stackelberg Solutions for Two-Person Games with Biased Information Patterns," IEEE Trans. Aut. Cont., Vol. AC-17, No. 6, Dec. 1972, pp. 791-798.

[6] Simaan, M. and Cruz, J.B., Jr., "On the Stackelberg Strategy in Nonzero-Sum Games," J.O.T.A., Vol. 11, No. 5, pp. 533-555, May 1973.

[7] Simaan, M. and Cruz, J.B., Jr., "Additional Aspects of the Stackelberg Strategy in Nonzero-Sum Games," J.O.T.A., Vol. 11, No. 6, June 1973, pp. 613-626.

[8] Simaan, M. and Cruz, J.B., Jr., "A Stackelberg Strategy for Games with Many Players," IEEE Trans. on Aut. Cont., Vol. AC-18, No. 3, June 1973, pp. 322-324.

[9] Simaan, M.,"Multicontrollers Strategies for Systems with Discrete Acquired and Biased Prior Information," Ph.D. Thesis, University of Illinois, Coordinated Science Lab. Report No.: R585, Sept. 1972.

[10] Sandell, N.R., Varaiya P., and Athans, M., "A Survey of Decentralized Control Methods for Large Scale Systems," Proc. ERDA Conference on System Engineering for Power: Status and Prospects, Henniker, N.H., Aug. 1975.

[11] Kokotovic, P.V., "Asymptotic Decoupling Method for High Order Differential Games," Proc. First Int. Conf. on the Th. and App. of Diff. Games, Univ. of Mass., Amherst, Mass., September 1969, pp. I.10.

[12] Medanic, J. and Andjelic, M., "On a Class of Differential Games Without Saddle-Point Solution," J.O.T.A., Vol. 8, No. 6, Dec. 1971, pp. 413-430. See also Ho, Y.C. "Comments on a paper by J. Medanic and M. Andjelic," J.O.T.A., Vol. 10, No. 3, pp. 187-189, September 1972, and Authors reply, pp. 190-191.

[13] Von Stackelberg, H., The Theory of the Market Economy, Oxford Univ. Press, Oxford, England 1952.

[14] Castanon, D. and Athans, M., "On Stochastic Dynamic Stackelberg Strategies," Automatica, Vol. 12, March 1976, pp. 177-183.

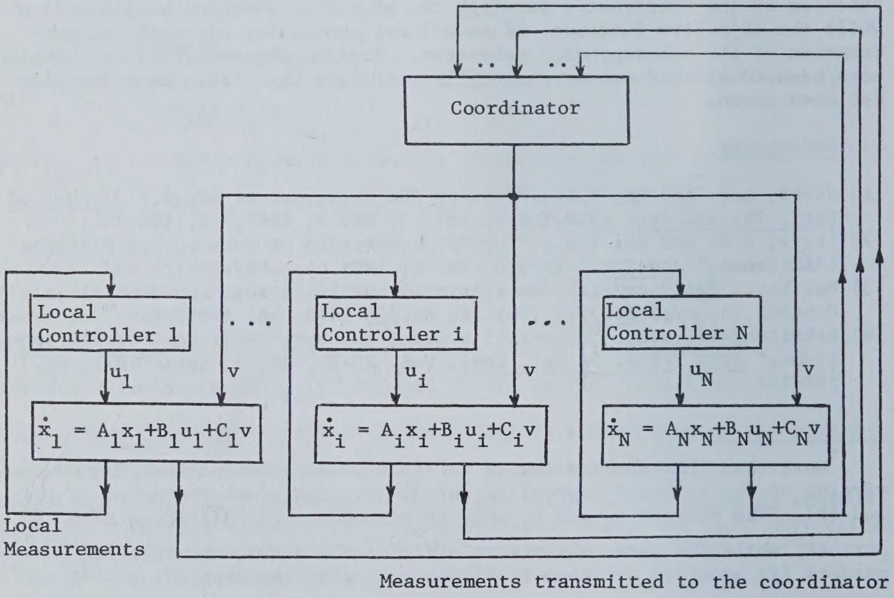

Coordinator
Local Controller 1
Local Controller i
Local Controller N
. . .
. . .
u_1
v
u_i
v
u_N
v
$\dot{x}_1 = A_1x_1+B_1u_1+C_1v$
$\dot{x}_i = A_ix_i+B_iu_i+C_iv$
$\dot{x}_N = A_Nx_N+B_Nu_N+C_Nv$
Local Measurements
Measurements transmitted to the coordinator

Figure 1: Two-Level System Configuration

ON THE MULTILEVEL STACKELBERG STRATEGIES IN LINEAR QUADRATIC SYSTEMS**

J. MEDANIC*
University of Illinois
Urbana, Illinois

D. RADOJEVIC
"Mihailo Pupin" Institute
Belgrade, Yugoslavia

ABSTRACT

Open-loop multilevel stackelberg strategies in deterministic sequential decision-making problems for continuous linear systems and quadratic criteria are developed. Characterization of the Stackelberg controls via the solution of a higher order square matrix Riccati differential equation is established, as well as the basic structural properties of the coefficient matrices of this differential equation from which the basic structural properties of its solution are inferred.

INTRODUCTION

The paper develops multilevel Stackelberg strategies in deterministic sequential decision-making problems for continuous linear systems and quadratic criteria. Basic results or two-level Stackelberg problems have been presented in [1] [2] [3] and surveyed in [4]; additional results have been presented in [5, 6, 7]. These references treat open-loop closed-loop and feedback (e.g., signaling free in [7]) strategies for both the continuous and discrete case and in both the deterministic and the stochastic setting. The results established so far give the necessary conditions for Stackelberg strategies for the general case, and characterize these strategies more completely for the class of linear-quadratic problems.

The general multilevel Stackelberg problem has been described in [8] and is now being studied. This paper shows that for linear-quadratic problems the results obtained for the two-level problem can be directly generalized to the multilevel case with an arbitrary number of sequential decisionmakers. For the sake of clarity results are presented in detail for the three level problem. This suffices to show how the generalized case with an arbitrary number of decisionmakers is solved.

The basic results established in the paper are the characterization of the Stackelberg controls through the solution of an $n \cdot 2^{k-1}$-th order square matrix Riccati differential equation, where k is the number of decisionmakers and n is the dimension of the dynamic system they control, the establishment of the basic structural properties of the coefficient matrices of that Riccati matrix differential equation, and from that, the inference of the basic structural properties of the solution.

PROBLEM FORMULATION AND DEVELOPMENT OF NECESSARY CONDITIONS

Consider an interconnected linear dynamic system in which three subsystems may be distinguished. Each subsystem is characterized by the ability to excert partial control over the system by choice of its control

---

* On leave from the "Mihailo Pupin" Institute, Belgrade, Yugoslavia.

** This work was supported in part by the Energy Research and Development Administration Contract No. ERDA E(49-18)-2088.

u_i and by a performance criterion $J_i(u_1,u_2,u_3,x_o)$ through which system performance is evaluated. It is assumed that a decision-making sequence is defined and that once a decisionmaker selects a strategy all decisionmakers following him in the decision-making sequence become aware of the selected strategy. For this class of control problems the Stackelberg strategy is the accepted solution concept.

The sequence $\{N_3,N_2,N_1\}$ implies that subsystem N_3 is the leader and selects its strategy first; N_2 is the first follower and selects its strategy second and N_1 is the second follower and selects its strategy last. Consequently, in making his decision N_1 knows the controls u_2 and u_3, N_2 must take into account the reaction of N_1 but knows u_3 and N_1 must take into account the sequential reactions of both N_2 and N_1. The simplest problem is solved by N_1 (an optimal control problem), a more complex problem by N_2 (a two-level Stackelberg problem) and the most complex by N_3 (a three-level Stackelberg problem). The complete solution of the problem is obtained by the solution of the leader's control problem, since the leader must solve problems faced by both N_1 and N_2 to determine their reactions to a given u_3 in order to select that control which is best with respect to J_3 taking into account the reactions of the followers. The leader's control problem is solved by deducing sequentially, in the order $\{N_1,N_2,N_3\}$, retrogressive with respect to the decision-making sequence, the Stackelberg controls for the subsystem. It will be seen that the effect of this procedure is that as each new control is characterized the dimensionality of the associated system, that serves as a differential constraint in the determination of the optimal strategy for the next control in the sequence, is doubled.

Let the interconnected system be defined by the linear differential equation

$$\dot{x} = Ax + B_1u_1 + B_2u_2 + B_3u_3 \tag{1}$$

and let the performance criteria associated with each subsystem be quaratic

$$J_i(u_1,u_2,u_3,x_o) = \frac{1}{2}\int_{t_o}^{T} (x^TQ_ix + \sum_{j=1}^{3} u_j^TR_{ij}u_j)dt + \frac{1}{2}\,x(T)^TF_ix(T), \tag{2}$$

where the usual positive (semi)definiteness conditions are imposed on A, B_i, Q_i, F_i and R_{ij}, i,j = 1,2,3, as in the associated optimal control problem. Consider now the problem of characterizing the open-loop Stackelberg strategies for the subsystems.

Consider the problem solved by N_1. Because N_1 is aware of the open-loop strategies selected by N_2 and N_3 the problem is reduced to an optimal control problem defined by (1) and (2), with i = 1, given u_2 and u_3, so that the necessary conditions take the form

$$\dot{x} = Ax + B_1u_1 + B_2u_2 + B_3u_3, \qquad x(t_o) = x_o \tag{3a}$$

$$\dot{p} = -Q_1x - A^Tp_1, \qquad p_1(T) = F_1x(T) \tag{3b}$$

$$0 = R_{11}u_1 + B_1^Tp_1 \tag{3c}$$

from which the control u_1 may be expressed as

$$u_1 = - R_{11}^{-1}B_1^Tp_1 \tag{4}$$

and is an implicit function of the controls u_2 and u_3, since they influence the costate vector p.

Consider now the problem solved by N_2. In deciding on the control u_2, N_2 must take into account the reaction of N_1 to a given u_2. The simplest way of characterizing the reaction of N_1 to u_2 is to substitute u_1 as given by (4) into (3a) to obtain the resulting two-point boundary value problem:

$$\dot{x} = Ax - S_1p_1 + B_2u_2 + B_3u_3, \quad S_1 = B_1R_{11}^{-1}B_1^T, \quad x(t_o) = x_o \tag{5a}$$

$$\dot{p} = - Q_1x - A^Tp_1, \qquad p_1(T) = F_1x(T) \tag{5b}$$

It may be viewed that (5b) characterizes the reaction of N_1 on the strategies selected by N_2 and N_3 while (5a) characterizes the dynamics of the interconnected system taking this reaction into account. Also, substituting (4) into (2) for i = 2 the performance criterion for N_2 taking into account the reaction of N_1 to a given u_2 and u_3 takes the form:

$$J_2 = \frac{1}{2}\int_{t_o}^{T} (x^TQ_2x + p_1^TS_{21}p_1 + u_2^TR_{22}u_2 + u_3^TR_{23}u_3)dt + \frac{1}{2}\, x(T)^TF_2x(T) \tag{6}$$

where $S_{21} = B_1R_{11}^{-1}R_{21}R_{11}^{-1}B_1^T$, so that the necessary conditions take the form:

$$\dot{x} = Ax - S_1p_1 + B_2u_2 + B_3u_3\ , \qquad x(t_o) = x_o \tag{7a}$$

$$\dot{p}_1 = -Q_1x - A^Tp_1 , \qquad p_1(T) = F_1x(T) \tag{7b}$$

$$\dot{p}_2 = -Q_2x - A^Tp_2 + Q_1n_1, \qquad p_2(T) = F_2x(T) - F_1n_1(T) \tag{7c}$$

$$\dot{n}_1 = -S_{21}p_1 + S_1p_2 + An_1, \quad n_1(t_o) = 0 \tag{7d}$$

$$0 = R_{22}u_2 + B_2^Tp_2$$

from which the control u_2 may be expressed as

$$u_2 = -R_{22}^{-1}B_2^Tp_2 \tag{8}$$

Consider finally the problem solved by N_3. After substitution of the control u_2 given by (10) into (9a) the necessary conditions are reduced to

$$\dot{x} = Ax - S_1p_1 - S_2p_2 + B_3u_3 , \quad S_2 = B_2R_{22}^{-1}B_2^T, \; x(t_o) = x_0 \tag{9a}$$

$$\dot{p}_1 = -Q_1x - A^Tp_1, \qquad p_1(T) = F_1x(T) \tag{9b}$$

$$\dot{p}_2 = -Q_2x - A^Tp_2 + Q_1n_1, \quad p_2(T) = F_2x(T) - F_1n_1(T) \tag{9c}$$

$$\dot{n}_1 = - S_{21}p_1 + S_1p_2 + An_1 , \quad n_1(t_o) = 0 \tag{9d}$$

As in the previous case (9b) characterizes the reaction of N_1 to the strategy selected by N_2, (9c) and (9d) now characterize the reaction of N_2 to the strategy selected by N_3 and (9a) now characterizes the dynamics of the interconnected system taking into account both the reactions of N_2 and N_1. The performance criterion $J_3(u_1,u_2,u_3,x_o)$ with these reactions taken into account takes the form

$$J_3 = \int_{t_o}^{T} (x^TQ_3x + p_1^TS_{31}p_1 + p_2^TS_{32}p_2 + u_3^TR_{33}u_3)dt + \tfrac{1}{2}x(T)^TF_3x(T) \tag{10}$$

where $S_{3j} = B_jR_{jj}^{-1}R_{3j}R_{jj}^{-1}B_j^T$, $j = 1,2$ and the necessary conditions take the form

$$\dot{x} = Ax - S_1p_1 - S_2p_2 - S_3p_3, \qquad x(t_o) = x_o \tag{11a}$$

$$\dot{p}_1 = -Q_1x - A^Tp_1, \qquad p_1(T) = F_1x(T) \tag{11b}$$

$$\dot{p}_2 = -Q_2x - A^Tp_2 + Q_1n_1 \qquad p_2(T) = F_2x(T) - F_1n(T) \tag{11c}$$

$$\dot{n}_1 = -S_{21}p_1 + S_1p_2 + An_1 \qquad n_1(t_o) = 0 \tag{11d}$$

$$\dot{p}_3 = -Q_3x - A^Tp_3 + Q_1n_2 + Q_2n_3, \qquad p_3(T) = F_3x(T) - F_1n_2(T) - F_2n_3(T) \tag{11e}$$

$$\dot{n}_2 = -S_{31}p_1 + S_1p_3 + An_2 + S_{21}w, \qquad n_2(t_o) = 0 \tag{11f}$$

$$\dot{n}_3 = -S_{32}p_2 + S_2p_3 + An_3 - S_1w, \qquad n_3(t_o) = 0 \tag{11g}$$

$$\dot{w} = -Q_1n_3 - A^Tw, \qquad w(T) = F_1n_3(T) \tag{11h}$$

with

$$u_3 = -R_{33}^{-1}B_3^Tp_3 \tag{12}$$

Characterization of Optimal Strategies

Because (11) is a linear systems and the boundary conditions are linear we look for linear relations between the variables w, p_i, and n_i, $i = 1,2,3$ and the state x of the form

$$p_i = K_ix , \quad i = 1,2,3 \tag{13a}$$

$$n_i = P_ix , \quad i = 1,2,3 \tag{13b}$$

$$w = Wx \tag{13c}$$

It follows that W, K_i and P_i in order for (11) and (13) to be consistent must satisfy

$$\dot{K}_1 + K_1A + A^TK_1 - K_1S_1K_1 - K_1S_2K_2 - K_1S_3K_3 + Q_1 = 0 , \quad K_1(T) = F_1 \tag{14a}$$

$$\dot{K}_2 + K_2A + A^TK_2 - K_2S_1K_1 - K_2S_2K_2 - K_2S_3K_3 + Q_2 - Q_1P_1 = 0, \qquad K_2(T) = F_2 - F_1P_1(T) \tag{14b}$$

$$\dot{K}_3 + K_3A + A^TK_3 - K_3S_1K_1 - K_3S_2K_2 - K_3S_3K_3 + Q_3 - Q_1P_2 - Q_2P_3 = 0 \qquad K_3(T) = F_3 - F_1P_2(T) - F_2P_3(T) \tag{14c}$$

$$\dot{P}_1 + P_1A - AP_1 - P_1S_1K_1 - P_1S_2K_2 - P_1S_3K_3 + S_{21}K_1 - S_1K_2 = 0, \qquad P_1(t_o) = 0 \tag{14d}$$

$$\dot{P}_2 + P_2A - AP_2 - P_2S_1K_1 - P_2S_2K_2 - P_2S_3K_3 + S_{31}K_1 - S_1K_3 - S_{21}W = 0,$$

$$P_2(t_o) = 0 \qquad (14e)$$

$$\dot{P}_3 + P_3A - AP_3 - P_3S_1K_1 - P_3S_2K_2 - P_3S_3K_3 + S_{32}K_2 - S_2K_3 + S_1W = 0,$$

$$P_3(t_o) = 0 \qquad (14f)$$

$$\dot{W} + WA + A^TW - WS_1K_1 - WS_2K_2 - WS_3K_3 + Q_1P_3 = 0, \quad W(T) = P_3(T) \qquad (14g)$$

The solution of (14) will provide W, K_i and P_i and substitution of (13) into (4), (8) and (12) determines the open-loop Stackelberg strategies in the form

$$u_i = -R_{ii}^{-1} B_i K_i \phi(t,t_o)x_o, \qquad i = 1,2,3 \qquad (15)$$

where $\phi(t,t_o)$ is the fundamental matrix of the system (14a) with P_i, $i = 1,2,3$ expressed through (13):

$$\dot{x} = (A-S_1K_1-S_2K_2-S_3K_3)x, \; x(t_o) = x_o \qquad (16)$$

The solution of (14) is difficult because it reduces to the solution of a nonlinear two-point boundary value problem and effective methods for the solution of general two-point boundary value problems involve considerable numerical complexity. More important, there do not exist proven methods for the study of the associated infinite time problem. It is therefore of interest to transform the equations (14) to a better known form for which there exist reliable methods of solution. It will be shown that (14) can be transformed to the solution of a higher order (4n x 4n) dimensional Riccati matrix differential equation with a given terminal condition.

Note that in (14) initial conditions are given for x(t) and $n_i(t)$, $i = 1,2,3$, and terminal conditions are given for w(t) and $p_i(t)$, $i = 1,2,3$ and moreover that these terminal conditions are linear functions of x(t) and $n_i(t)$. Reorder, therefore the system (14) by grouping x(t) and $n_i(t)$ into the augmented vector $\tilde{x}(t)$ and w(t) and $p_i(t)$ into the augmented vector $\tilde{p}(t)$ and bring it to the form:

$$\left\| \begin{array}{c} \dot{x} \\ \dot{n}_1 \\ \dot{n}_2 \\ \dot{n}_3 \\ \hline \dot{p}_1 \\ \dot{p}_2 \\ \dot{p}_3 \\ \dot{w} \end{array} \right\| = \left\| \begin{array}{cccc|cccc} A & 0 & 0 & 0 & -S_1 & -S_2 & -S_3 & 0 \\ 0 & A & 0 & 0 & -S_{21} & S_1 & 0 & 0 \\ 0 & 0 & A & 0 & -S_{31} & 0 & S_1 & S_{21} \\ 0 & 0 & 0 & A & 0 & -S_{32} & S_2 & -S_1 \\ \hline -Q_1 & 0 & 0 & 0 & -A^T & 0 & 0 & 0 \\ -Q_2 & Q_1 & 0 & 0 & 0 & -A^T & 0 & 0 \\ -Q_3 & 0 & -Q_1 & -Q_2 & 0 & 0 & -A^T & 0 \\ 0 & 0 & 0 & Q_1 & 0 & 0 & 0 & -A^T \end{array} \right\| \left\| \begin{array}{c} x \\ n_1 \\ n_2 \\ n_3 \\ \hline p_1 \\ p_2 \\ p_3 \\ w \end{array} \right\|$$

Dashed lines have been used to delineate the equations that define the classical optimal control problem and the two-layer Stackelberg problem [3] respectively. The system can now be written in the compact form

$$\dot{\tilde{x}} = \tilde{A}\tilde{x} - \tilde{S}\tilde{p} \ , \quad \tilde{x}(t_o) = \tilde{x}_o \ , \quad \tilde{x}_o = |x_o^T \ 0 \ 0 \ 0|^T$$
$$\dot{\tilde{p}} = -\tilde{Q}\tilde{x} - \tilde{A}^T\tilde{p} \quad \tilde{p}(T) = \tilde{F}\tilde{x}(T) \tag{18}$$

where the augmented matrices $\tilde{A}$, $\tilde{S}$, $\tilde{Q}$ and $\tilde{F}$ are given by

$$\tilde{A} = \left\| \begin{array}{c|c} \hat{A} & 0 \\ \hline 0 & \hat{A} \end{array} \right\| = \left\| \begin{array}{cc|cc} A & 0 & 0 & 0 \\ 0 & A & 0 & 0 \\ \hline 0 & 0 & A & 0 \\ 0 & 0 & 0 & A \end{array} \right\| , \quad \tilde{S} = \left\| \begin{array}{c|c} \hat{S}_{11} & \hat{S}_{12} \\ \hline \hat{S}_{21} & \hat{S}_{22} \end{array} \right\| = \left\| \begin{array}{cc|cc} S_1 & S_2 & S_3 & 0 \\ S_{21} & -S_1 & 0 & 0 \\ \hline S_{31} & 0 & -S_1 & -S_{21} \\ 0 & S_{32} & -S_2 & S_1 \end{array} \right\|$$

$$\tilde{Q} = \left\| \begin{array}{c|c} \hat{Q}_{11} & 0 \\ \hline \hat{Q}_{21} & \hat{Q}_{22} \end{array} \right\| = \left\| \begin{array}{cc|cc} Q_1 & 0 & 0 & 0 \\ Q_2 & -Q_1 & 0 & 0 \\ \hline Q_3 & 0 & -Q_1 & -Q_2 \\ 0 & 0 & 0 & Q_1 \end{array} \right\| , \quad \tilde{F} = \left\| \begin{array}{c|c} \hat{F}_{11} & 0 \\ \hline \hat{F}_{21} & \hat{F}_{22} \end{array} \right\| = \left\| \begin{array}{cc|cc} F_1 & 0 & 0 & 0 \\ F_2 & -F_1 & 0 & 0 \\ \hline F_3 & 0 & -F_1 & -F_2 \\ 0 & 0 & 0 & F_1 \end{array} \right\| \tag{19}$$

Everywhere in the sequil "curled" variables will refer to 4n x 4n matrices, "hatted" variables to 2n x 2n matrices and bare variables to the original n x n matrices, where the relations between the various variables are uniquely defined by (19). System (18) is now solved in the usual way by defining the linear transformation

$$p(t) = K(t) \ x(t) \tag{20}$$

whereby it follows that K(t) must satisfy the Riccati equation

$$\dot{\tilde{K}} + \tilde{A}^T\tilde{K} + \tilde{K}\tilde{A} - \tilde{K}\tilde{S}\tilde{K} + \tilde{Q} = 0, \quad K(T) = F \tag{21}$$

If a solution of (21) exists, and in particular if there exists the solution of the associated algebraic Riccati equation for the infinite time problem this method of characterization is more expedient despite the increased order of the Riccati matrix since well known algorithms for its solution have been developed. Furthermore, this characterization allows the powerful algebraic methods developed for the algebraic Riccati equation [6] [7] of the optimal control problem to be brought into the Stackelbert solutions. However, the usual conditions of positive (semi) definiteness and symmetry assumed there do not apply here and the problem must be studied separately.

Note finally that if the solution to (21) exists and is represented in the form

$$\tilde{K} = \left\|\begin{array}{c|c} \hat{K}_{11} & \hat{K}_{12} \\ \hline \hat{K}_{21} & \hat{K}_{22} \end{array}\right\| = \left\|\begin{array}{cc|cc} K_{11} & K_{12} & K_{13} & K_{14} \\ K_{21} & K_{22} & K_{23} & K_{24} \\ \hline K_{31} & K_{32} & K_{33} & K_{34} \\ K_{41} & K_{42} & K_{43} & K_{44} \end{array}\right\| \qquad (22)$$

we have from (15) and (24)

$$\begin{aligned} K_1 &= K_{11} + K_{12}P_1 + K_{13}P_2 + K_{14}P_3 \\ K_2 &= K_{21} + K_{22}P_1 + K_{23}P_2 + K_{24}P_3 \\ K_3 &= K_{31} + K_{32}P_1 + K_{33}P_2 + K_{34}P_3 \\ W &= K_{41} + K_{42}P_1 + K_{43}P_2 + K_{44}P_3 \end{aligned} \qquad (23)$$

Substituting K_i, $i = 1,2,3$ and W from (25) into (16d) - (16f) and solving for P_i, $i = 1,2,3$, from $t = t_o$ on obtains the optimal controls (17) with K_i given by (25).

Structural Properties of the Associated Riccati Matrix Differential Equation

The open-loop solution of the multilevel Stackelberg controls is intrinsically related to the Riccati matrix differential equation (21). It is immediately evident in this case that the coefficient matrices $\{\tilde{A},\tilde{S},\tilde{Q},\tilde{F}\}$ do not possess the symmetricity and semi-positive definiteness properties present in the optimal control problem. For that reason none of the properties of the solution K, present in that much studied case can be assumed to hold, including the existence of the solution on a given time interval $[t_o,T]$. The question of a finite blow-up time is at present open, save for the conservative bounds that can be placed on it from the general norm considerations related to the growth of the solution of the Riccati differential equations. But, assuming that in a particular case the solution does exist on a time interval $[t_o,T]$ it is possible to derive some of the structural properties of the solution. Recalling the defining relations (19) and (22) for the various matrices call the quadruple $\{\tilde{A},\tilde{S},\tilde{Q},\tilde{F}\}$ the coefficient matrices of the Riccati matrix differential equation. Note the following properties of the various matrices:

(a) all basic matrices defined in (9) are symmetric, except the matrix A

$$\begin{aligned} &\text{(b)}\quad \hat{S}_{ij}^{T} = \hat{S}_{ij},\ \hat{S}_{ii} = -\hat{S}_{jj}^{T}\quad,\ i,j = 1,2,\quad i \neq j \\ &\text{(c)}\quad \hat{Q}_{ij}^{T} = \hat{Q}_{ij},\ \hat{Q}_{ii} = -\hat{Q}_{jj}^{T}\quad,\ i,j = 1,2,\quad i \neq j \\ &\text{(d)}\quad \hat{F}_{ij}^{T} = \hat{F}_{ij},\ \hat{F}_{ii} = -\hat{F}_{jj}^{T}\quad,\ i,j = 1,2,\quad i \neq j \end{aligned} \qquad (24)$$

It is possible to state the following result:

Theorem 1: The solution $\tilde{K}(t)$ of the associated Riccati matrix differential equation (8) for coefficient matrices $\{\tilde{A},\tilde{S},\tilde{Q},\tilde{F}\}$ satisfying conditions (b) - (d) has the structure characterized by the properties

$$\hat{K}_{ij}^{T} = \hat{K}_{ij} \tag{25}$$

$$\hat{K}_{ii}^{T} = \hat{K}_{jj}$$

for i,j = 1,2 and i ≠ j.

The proof of this theorem relies on the uniqueness theorem of differential equations, if a solution to the considered equation on the finite interval $[t_o,T]$ exists. Assuming the solution exists the proof proceeds by showing that the solution, which then is unique, possesses properties (13). To that effect decompose $\tilde{A},\tilde{S},\tilde{Q},\tilde{F}$ and $\tilde{K}$ into 2n x 2n matrices, i.e., utilize the hatted variables to write (8) as a system of 4 coupled matrix differential equations:

$$\dot{\hat{K}}_{11} + \hat{A}^T\hat{K}_{11} + \hat{K}_{11}\hat{A} - \hat{K}_{11}(\hat{S}_{11}\hat{K}_{11}+\hat{S}_{12}\hat{K}_{21}) - \hat{K}_{12}(\hat{S}_{21}\hat{K}_{11}-\hat{S}_{11}^T\hat{K}_{21}) + \hat{Q}_{11} = 0, \quad \hat{K}_{11}(T) = \hat{F}_{11} \tag{26a}$$

$$\dot{\hat{K}}_{12} + \hat{A}^T\hat{K}_{12} + \hat{K}_{12}\hat{A} - \hat{K}_{11}(\hat{S}_{11}\hat{K}_{12}+\hat{S}_{12}\hat{K}_{22}) - \hat{K}_{12}(\hat{S}_{21}\hat{K}_{12}-\hat{S}_{11}^T\hat{K}_{22}) = 0, \quad \hat{K}_{12}(T) = \hat{0} \tag{26b}$$

$$\dot{\hat{K}}_{21} + \hat{A}^T\hat{K}_{21} + \hat{K}_{21}\hat{A} - \hat{K}_{21}(\hat{S}_{11}\hat{K}_{11}+\hat{S}_{12}\hat{K}_{21}) - \hat{K}_{22}(\hat{S}_{21}\hat{K}_{11}-\hat{S}_{11}^T\hat{K}_{21}) + \hat{Q}_{21} = \hat{0}, \quad \hat{K}_{21}(T) = \hat{F}_{21} \tag{26c}$$

$$\dot{\hat{K}}_{22} + \hat{A}^T\hat{K}_{22} + \hat{K}_{22}\hat{A} - \hat{K}_{21}(\hat{S}_{11}\hat{K}_{12}+\hat{S}_{12}\hat{K}_{22}) - \hat{K}_{22}(\hat{S}_{21}\hat{K}_{12}-\hat{S}_{11}^T\hat{K}_{22}) - \hat{Q}_{11}^T = \hat{0}, \quad \hat{K}_{22}(T) = - F_{11}^T \tag{26d}$$

And clearly if K is to enjoy properties (13) $\hat{K}_{12}^T$ and $\hat{K}_{21}^T$ must satisfy the same differential equations as $\hat{K}_{12}$ and $\hat{K}_{21}$, respectively, while the negative transpose of the matrix differential equation for $\hat{K}_{22}$ must reduce to the same exuation satisfied by $\hat{K}_{11}$. Suppose therefore that (13) holds and consider whether it then satisfied (14). Transpose and premultiply by a negative sign the equation for $\hat{K}_{22}$ with the result

$$-\dot{\hat{K}}_{22}^T + (-\hat{K}_{22}^T)\hat{A} + \hat{A}^T(-\hat{K}_{22}^T) - [(-\hat{K}_{22}^T)\hat{S}_{12} + \hat{K}_{12}^T\hat{S}_{11}^T]\hat{K}_{21} - [(-\hat{K}_{22}^T)\hat{S}_{11}+\hat{K}_{12}^T\hat{S}_{21}](-\hat{K}_{22}^T) + \hat{Q}_{11} = 0, \quad -\hat{K}_{22}^T(T) = \hat{F}_{11} \tag{27}$$

or after rearranging the terms

$$-\dot{\hat{K}}_{22}^T + \hat{A}^T(-\hat{K}_{22}^T) + (-\hat{K}_{22}^T)\hat{A} - (-\hat{K}_{22}^T)(\hat{S}_{11}(-\hat{K}_{22}^T) + \hat{S}_{12}\hat{K}_{21}^T) - \hat{K}_{12}^T(\hat{S}_{21}(-\hat{K}_{22}^T) - \hat{S}_{11}^T\hat{K}_{21}^T) + \hat{Q}_{11} = 0 \tag{28}$$

Clearly, $-K^T_{22}$ will satisfy the same matrix differential equation as $\hat{K}_{11}$, see (14), if $\hat{K}^T_{12} = \hat{K}_{12}$ and $\hat{K}^T_{21} = \hat{K}_{21}$. Thus, symmetricity of $\hat{K}_{12}$ and $\hat{K}_{21}$, which has been assumed, is sufficient to guarantee that $\hat{K}^T_{22} = -\hat{K}_{11}$. On the other hand we show that the antisymmetry of $\hat{K}_{11}$ and $\hat{K}_{22}$ is sufficient to guarantee that the symmetry of $\hat{K}_{ij}$, i = j holds. Transposing the equation for $\hat{K}_{12}$ one obtains

$$\dot{\hat{K}}^T_{12} + \hat{K}^T_{12}(\hat{A}-\hat{S}^T_{11}\hat{K}^T_{11}) + (\hat{A}^T+\hat{K}^T_{22}\hat{S}_{11})\hat{K}^T_{12} - \hat{K}^T_{12}\hat{S}_{21}\hat{K}^T_{21} - \hat{K}^T_{22}\hat{S}_{12}\hat{K}^T_{11} = 0,$$
$$\hat{K}^T_{12}(T) = 0 \qquad (29)$$

and rearranging terms and utilizing the mutual antisymmetry of $\hat{K}_{11}$ and $\hat{K}_{22}$ we obtain

$$\dot{\hat{K}}^T_{12} + (\hat{A}-\hat{K}_{11}\hat{S}_{11})\hat{K}^T_{12} + \hat{K}^T_{12}(\hat{A}+\hat{S}^T_{11}\hat{K}_{22}) - \hat{K}^T_{12}\hat{S}_{21}\hat{K}^T_{21} - \hat{K}_{11}\hat{S}_{12}\hat{K}_{22} = 0,$$
$$\hat{K}^T_{12}(T) = 0 \qquad (30)$$

and therefore $\hat{K}^T_{12}$ satisfies the same matrix differential equation as $\hat{K}_{12}$, with the same terminal condition, which proves its symmetry. The symmetry of $\hat{K}_{21}$ is shown similarly, which completes the proof.

Theorem 2: If a solution $\tilde{K}$ exists on the interval [to,T] the submatrices $\hat{K}_{ij}$, i,j = 1,2, i ≠ j are positive semidefinite and monotone nonincreasing functions of time.

Proof: Consider first $\hat{K}_{21}$. Assuming that $\hat{K}_{11}$ to be the solution of (26a) and taking into account that $-K_{22} = K^T_{11}$ the differential equation for $\hat{K}_{21}$ may be written in the form

$$\dot{\hat{K}}_{21} + (\hat{A}-\hat{S}_{11}\hat{K}_{11})^T\hat{K}_{21} + \hat{K}_{21}(\hat{A}-\hat{S}_{11}\hat{K}_{11}) - \hat{K}_{21}\hat{S}_{12}\hat{K}_{21} + \hat{Q}_{21} + \hat{K}_{11}\hat{S}_{21}\hat{K}_{11} = 0$$
$$\hat{K}_{21}(T) = \hat{F}_{21}$$

Because $\hat{F}_{12} \geq 0$, $\hat{Q}_{21} + \hat{K}_{11}{}^T\hat{S}_{21}\hat{K}_{11} \geq 0$, $\hat{S}_{12} \geq 0$ and because $\hat{F}_{12}$, $\hat{Q}_{21}$, $\hat{S}_{21}$ and $\hat{S}_{12}$ are symmetric it is immediately clear that $\hat{K}_{21}$ enjoys all the well known properties of the optimal gain matrix of an associated linear regulator problem, and in particular positive semidefiniteness and monotone nonincreasing character. Similar conclusion is drawn for $\hat{K}_{12}$ by observing that it satisfies the equation

$$\dot{\hat{K}}_{12} + (\hat{A}-\hat{S}^T_{11}\hat{K}^T_{11})^T\hat{K}_{12} + \hat{K}_{12}(\hat{A}-\hat{S}^T_{11}\hat{K}^T_{11}) - \hat{K}_{12}\hat{S}_{21}\hat{K}_{12} + \hat{K}_{11}\hat{S}_{12}\hat{K}_{11}\hat{S}_{13}\hat{K}^T_{11} = 0$$
$$K_{12}(T) = 0$$

and again the characteristic properties of $\hat{S}_{12}$ and $\hat{S}_{21}$ are sufficient to establish the properties of $\hat{K}_{12}$ cited in the theorem.

As a consequence, the burden of computing the exact solution $\tilde{K}(t)$ is precisely equal to the burden associated with solving the standard Riccati equation of order $4n \times 4n$ with symmetrical coefficient matrices and involves the solution of a system of $8n^2 + 2n$ differential equations. If it is assumed that the computation time increases with the power of 3 with the number of equations to be solved simultaneously then for a direct solution one uses up time proportional to $(8n^2+2n)^3 + (3n^2)^3$, where the second term comes in due to the necessity of solving for P_i, $i = 1,2,3$; on the other hand if K iterations are needed for satisfactory convergence to the solution of the two-point boundary value problem (4) - (6) then the total time used in that case is proportional to $K[4n^2)^3+(3n^2)^3]$. Therefore, for a one dimensional system (n=1) the Riccati method would be superior if $K > 10$, and as k the dimensionality of the system increases, it would be superior in the limit for $K > 6$.

Conclusion

The paper develops explicit expressions for multilayer open-loop Stackelberg strategies for continuous linear-quadratic sequential decision-making problems. It is shown that the two-layer solution [3] may directly be generalized to multilayer problems. It is shown, in particular that the problem can always be reduced to the characterization by an associated Riccati solution. The order of the associated Riccati equation is $n \cdot 2^{k-1}$ where n is the order of the system and k is the number of sequential decisionmakers.

References

1. H. von Stackelberg, The Theory of the Market Economy, Oxford University Press, Oxford, England, 1952.
2. C. I. Chen and J. B. Cruz, Jr., "Stackelberg Solution for Two-Person Games with Biased Information Patterns," IEEE Trans. on Aut. Control, Vol. AC-17, No. 6, 1972.
3. M. Simaan and J. B. Cruz, Jr., "On the Stackelberg Strategy in Nonzero-Sum Games," J. Opt. Theory and Appl., Vol. 11, No. 5, 1973.
4. J. B. Cruz, Jr., "Survey of Nash and Stackelberg Equilibrium Strategies in Dynamic Games," Ann. of Economic and Social Measurement, Vol. 4, No. 2, 1975.
5. M. Simaan and J. B. Cruz, Jr., "Additional Aspects of the Stackelberg Strategy in Nonzero-Sum Games," JOTA, Vol. 11, No. 6, June 1973.
6. T. Basar, "On the Relative Leadership Property of Stackelberg Strategies," JOTA, Vol. 11, No. 6, June 1973.
7. D. Castanon and M. Athans, "On the Stochastic Dynamic Stackelberg Strategies," Proc. VI World Congress of IFAC, Boston, Aug. 1975, ISA.
8. J. B. Cruz, Jr., "Stackelberg Strategies for Multilevel Systems," in Directions in Large-Scale Systems, ed. by Y. C. Ho and S. K. Mitter, Plenum Press, 1975.

A NEW APPROACH TO OUTPUT FEEDBACK STABILIZATION*

ADLY T. FAM
School of Engineering
University of California
Irvine, California 92717

JAMES S. MEDITCH
Department of Electrical Engineering
University of Washington
Seattle, Washington 98195

ABSTRACT

The stabilizability of linear, time-invariant, discrete-time systems by constant gain output feedback is formulated as a geometric problem in the "companion parameter space". The properties of the stability domain in this parameter space are investigated. This leads, in addition to characterizing the class of systems stabilizable by output feedback, to some interesting results related to the robustness of output feedback, and to an application of integral geometry which shows the limitations of output feedback as a means for stabilizing linear discrete-time systems. The geometric approach utilized here provides efficient computational techniques for both the recursive construction of the stability domain for systems of any order and solution of the problem of stabilizability of linear systems by output feedback.

I. INTRODUCTION

The work presented in this paper is motivated by a fundamental unsolved problem in feedback control, the problem of stabilizability of linear systems by constant gain output feedback. We are concerned with systems of the form

$$x(t+1) = Ax(t) + Bu(t), \qquad y(t) = Cx(t) \tag{1}$$

where $x(\cdot)$, $u(\cdot)$ and $y(\cdot)$ are vectors of n, r and m components, respectively; A, B and C are $n \times n$, $n \times r$ and $m \times n$ matrices, respectively, and $t = 0, 1, \ldots$ denotes discrete time. All vectors and matrices are real. We assume that (A, B) is controllable and (A, C) is observable.

Under output feedback of the form

$$u(t) = Ky(t) \tag{2}$$

where K is an $r \times m$ matrix, the system (1) becomes

$$x(t+1) = (A + BKC)\, x(t) \tag{3}$$

To decide whether there exists a matrix K such that (A + BKC) is asymptotically stable, and to find such a K if it exists, where A, B and C are given and A is unstable, is a nonlinear parametric eigenvalue problem. Many investigations in the current literature offer solutions that either replace this problem with a more difficult one or solve a restricted version of it. Examples of each are the decision methods investigated in [1] and the pole placement by output feedback considered in [2] , respectively. The work in [2] leads to a sufficient condition for stabilizing linear systems by output feedback.

In this work, we first consider the case of single-input/single-output systems. We develop a geometric approach to solve the stabilizability problem for this case, and then show how it can be extended directly to both the single-input/multi-output and the multi-input/single-output cases. The general case of multi-input/multi-output systems will be considered in [3].

To begin, consider the single-input/single-output case of (1) and assume that A is in the companion canonical form, obtained by transforming

---

*Research supported in part by the National Science Foundation under Grant No. ENG73-03729 A02.

the system to the companion controllable canonical form

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -\alpha_n & -\alpha_{n-1} & -\alpha_{n-2} & \cdots & -\alpha_2 & -\alpha_1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix} \quad \text{and}$$

$C = -[\beta_n \ \beta_{n-1} \ldots \beta_1] = -\underline{\beta}$. Under the scalar output feedback $u = Ky$, the last row of A is replaced by $-[(\alpha_n + K\beta_n) \ldots (\alpha_1 + K\beta_1)] = -(\underline{\alpha} + K\underline{\beta})$ where $\underline{\alpha} \triangleq (\alpha_n \, \alpha_{n-1} \ldots \alpha_1)$ and $\underline{\beta} \triangleq (\beta_n \, \beta_{n-1} \ldots \beta_1)$. The characteristic equation of A is

$$P(z) = z^n + \alpha_1 z^{n-1} + \alpha_2 z^{n-2} + \ldots + \alpha_n = 0 \tag{4}$$

The zeros of the polynomial P(z) are the eigenvalues of A and are determined uniquely by $\underline{\alpha}$. The vector $\underline{\alpha}$ could be represented as a point in the n-dimensional Euclidean space E^n with coordinates $(\alpha_n, \alpha_{n-1}, \ldots, \alpha_1)$. Application of the output feedback $u = Ky$ results in adding the vector $K\underline{\beta}$ to $\underline{\alpha}$ in E^n. Since both $\underline{\alpha}$ and $\underline{\beta}$ are fixed for a given system, the locus of the point $\underline{\alpha} + K\underline{\beta}$ is clearly a straight line. Every point on that straight line corresponds to one value of K as shown in Fig. 1. The fact that the locus of $\underline{\alpha} + K\underline{\beta}$ is a straight line (a hyperplane for the multi-output case) is fundamental to the development of this geometric approach. Since there is a one-to-one correspondence between each polynomial P(z) and its companion matrix A in the form given above and a point $\underline{\alpha}$ in E^n, we call the space E^n of such $\underline{\alpha}$'s the "companion parameter space". The collection of points $\underline{\alpha}$ in E^n corresponding to a polynomial P(z) with zeros in the open unit disc corresponds to the set of stable systems and is called the stability domain $\mathscr{D}_n$. The stabilizability problem reduces now to establishing whether or not $\underline{\alpha} + K\underline{\beta}$ intersects $\mathscr{D}_n$ as indicated in Fig. 1.

In Section II we investigate properties of the stability domain $\mathscr{D}_n$, some of which suggest efficient computation techniques for its recursive construction and for characterizing its convex hull. Section III is devoted to formulating stabilizability by output feedback as a geometric problem in E^n, and indicating relations to robustness of output feedback. An application of integral geometry, illustrated by some examples that show the limitations of output feedback, is presented in Section IV. Our conclusion is given in Section V.

II. THE STABILITY DOMAIN

In this section, we shall, in turn, consider the nature of the boundary of $\mathscr{D}_n$, show how the domain can be constructed recursively for successive n, and characterize its convex hull. The latter constitutes the central result of this section and presented as a theorem.

Every point on the boundary $\partial\mathscr{D}_n$, corresponds to a polynomial with at least one zero on the unit circle (either one zero at +1 or at -1, or a conjugate pair with magnitude 1). Hence, every $\underline{\alpha} \in \partial\mathscr{D}_n$ is associated with a polynomial $P_n(z)$ such that

$$P_n(e^{i\theta}) = e^{in\theta} + \sum_{j=1}^{n} \alpha_j \, e^{i(n-j)\theta} = 0 \text{ for some } \theta \in [0, 2\pi) \text{ where}$$

$i = \sqrt{-1}$. The three cases $z = -1$, $z = 1$ and $z = e^{i\theta}$, $\theta \in (0, 2\pi)$ produce the following three hypersurfaces, respectively:

$$(-1)^n + \alpha_1(-1)^{n-1} + \ldots + \alpha_n = 0 \tag{5}$$

$$1 + \alpha_1 + \alpha_2 + \ldots + \alpha_n = 0 \tag{6}$$

$$\left.\begin{array}{l}\cos n\theta + \alpha_1 \cos((n-1)\theta) + \ldots + \alpha_n = 0 \\ \sin n\theta + \alpha_1 \sin((n-1)\theta) + \ldots + \alpha_{n-1} \sin\theta = 0\end{array}\right\}, \ \theta \ (0, 2\pi) \qquad (7)$$

Equations (5) and (6) are those of two (n-1)-dimensional hyperplanes while (7) is that of a hypersurface that is produced by the movement of an (n-2)-dimensional hyperplane. Every value of θ in (7) corresponds to an (n-2)-dimensional hyperplane contained in the hypersurface (7).

The domain containing the origin and bounded by the surfaces (5), (6) and (7) is the stability domain $\mathscr{D}_n$.

Example 1. The determination of $\mathscr{D}_1$ and $\mathscr{D}_2$, both of which we shall need in the sequel, is easily carried out via direct application of (5), (6) and (7). First, $\partial\mathscr{D}_1$ is defined by

$$(-1)^1 + \alpha_1 = 0$$

or $\alpha_1 = 1$, and by

$$1 + \alpha_1 = 0$$

or $\alpha_1 = -1$. Hence, we have immediately that $\mathscr{D}_1$ is given by $-1 < \alpha_1 < 1$.

Similarly, $\partial\mathscr{D}_2$ is defined by the relations

$$(-1)^2 + \alpha_1(-1)^1 + \alpha_2 = 0 \qquad (8)$$

$$1 + \alpha_1 + \alpha_2 = 0 \qquad (9)$$

and

$$\cos 2\theta + \alpha_1 \cos\theta + \alpha_2 = 0 \qquad (10)$$

$$\sin 2\theta + \alpha_1 \sin\theta = 0 \qquad (11)$$

The use of trigonometric identities reduces (10) and (11) to the single relation

$$\alpha_2 = 1 \qquad (12)$$

Equations (8), (9) and (12) determine the boundaries of the triangle shown in Fig. 2. The interior of this triangle is $\mathscr{D}_2$. □

We are now in a position to show how the $\mathscr{D}_n$ can be constructed recursively from $\mathscr{D}_1$ and $\mathscr{D}_2$.

Constructing $\mathscr{D}_n$ recursively

Let $\left[\alpha_1^j \ . \ . \ . \ \alpha_j^j\right]'$ where prime denotes the transpose represent a point in $\mathscr{D}_j$. Then from the properties of multiplication of polynomials and the fact that the product of two stable polynomials is stable, it can be shown that

$$\begin{bmatrix} 1 \\ \alpha_1^{n+1} \\ \cdot \\ \cdot \\ \cdot \\ \alpha_{n+1}^{n+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & & \bigcirc \\ \alpha_1^1 & 1 & 0 & & \\ 0 & \alpha_1^1 & 1 & & \\ & & \ddots & \ddots & 1 \\ \bigcirc & & & \ddots & \alpha_1^1 \end{bmatrix} \begin{bmatrix} 1 \\ \alpha_1^n \\ \alpha_2^n \\ \cdot \\ \cdot \\ \alpha_n^n \end{bmatrix},$$

$$\begin{bmatrix} 1 \\ \alpha_1^{n+2} \\ \cdot \\ \cdot \\ \cdot \\ \alpha_{n+2}^{n+2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & & & \bigcirc \\ \alpha_1^2 & 1 & 0 & & \\ \alpha_2^2 & \alpha_1^2 & 1 & \ddots & \\ 0 & \alpha_2^2 & \ddots & \ddots & 1 \\ & \ddots & \ddots & \ddots & \alpha_1^2 \\ \bigcirc & & & \ddots & \alpha_2^2 \end{bmatrix} \begin{bmatrix} 1 \\ \alpha_1^n \\ \cdot \\ \cdot \\ \cdot \\ \alpha_n^n \end{bmatrix} ; \text{ n even integer.}$$

Hence, from a knowledge of $\mathscr{D}_1$ and $\mathscr{D}_2$, any $\mathscr{D}_n$ can be constructed. For the sake of illustration, the construction of $\mathscr{D}_3$ from $\mathscr{D}_1$ and $\mathscr{D}_2$ is depicted in Fig. 3. The same recursive relations apply to the boundary of the stability domain or for generating domains that satisfy any property invariant under polynomial multiplication. For example, stability domains which are restricted to part of $\mathscr{D}_n$ that corresponds to the zeros of P(z) lying in a region of the complex plane other than the open unit disc, say for the purpose of achieving robustness or imposing tolerances, can be constructed recursively using the same relations.

Convex hull of the stability domain

The next theorem characterizes the convex hull of the stability domain. The simplicity by which the convex hull is obtained is of value in finding the best necessary condition of stabilizability by output feedback that is computable by linear inequalities.

Theorem 1

The convex hull of $\mathscr{D}_n$ is a polyhedron whose vertices correspond to all polynomials $P_n(z)$ with zeros in the set {1, -1}.

Theorem 1 states that by substituting for the n-tuple $(\lambda_n, \lambda_2, \ldots, \lambda_1)$ in $P_n(z) = \prod_{i=1}^{n} (z - \lambda_i) = z^n + \sum_{j=1}^{n} z^{n-j} \alpha_j$ the values (1, 1, ..., 1), (1, 1, ..., -1), ..., (-1, -1, ..., -1), we get n+1 vectors $\underline{\alpha}$ representing n+1 linearly independent points in E^n. These n+1 points define uniquely a geometric n-simplex with the n+1 points as vertices.

Proof (By induction)

Theorem 1 is true for $\mathscr{D}_1$ and $\mathscr{D}_2$ by direct verification (see Example 1). Assume the theorem is true for $\mathscr{D}_j$, $1 \leq j \leq m$. This assumption implies that any polynomial $P_j(z)$, $1 \leq j \leq m$ is expressible as a linear combination of the polynomials corresponding to the vertices specified by the theorem. Hence, we have

$$P_j(z) = \sum_{i=0}^{i} a_i \, \Phi_{ji}(z), \qquad 1 \leq j \leq m$$

where

$$a_i \geq 0, \quad \sum_{i=0}^{j} a_i = 1$$

and

$$\Phi_{ji} = (z-1)^{j-i} \, (z+1)^i$$

Assume m is even (if not, consider $\left\{\mathscr{D}_j\right\}$, $1 \leq j \leq m-1$). Since both $P_2(z)$ and $P_m(a)$ have real coefficients, and $P_{m+2}(z) = P_2(z) \, P_m(z)$, then

$$P_{m+2}(z) = \left(\sum_{\ell=0}^{2} b_\ell \Phi_{2\ell}(z)\right) \left(\sum_{i=0}^{m} a_i \Phi_{mi}(z)\right) \tag{13}$$

and since

$$\sum_{\ell=0}^{2} b_\ell = \sum_{i=0}^{m} a_i = 1$$

then

$$\left(\sum_{\ell=0}^{2} b_\ell\right) \left(\sum_{i=0}^{m} a_i\right) = 1 \tag{14}$$

Indices of Φ change under multiplication as follows:

$$\begin{aligned}\Phi_{2\ell}\,\Phi_{mi} &= (z-1)^{2-\ell}\,(z+1)^{\ell}\,(z-1)^{m-i}\,(z+1)^{i} \\ &= (z-1)^{(2+m)-(\ell+i)}\,(z+1)^{(\ell+i)} = \Phi_{(2+m),(\ell+i)}\end{aligned} \tag{15}$$

From equations (13), (14) and (15), the theorem holds for m+2 and, hence, by induction, for all even positive integers. But since it holds for $\mathscr{D}_1$ and all $\mathscr{D}_m$ where m is even, and since $P_{m+1} = P_1 P_m$, also for m even, and repeating the same argument as before, then Theorem 1 is true for all $\mathscr{D}_m$, $1 \leq m$. □

Theorem 1 shows that the convex hull of $\mathscr{D}_m$ is very easy to obtain. This is important, since it enables us to obtain the best necessary condition for stabilizability of linear systems by output feedback that is expressible in the form of linear inequalities. This also suggests interesting computational techniques.

Example 2. Convex hull of $\mathscr{D}_3$: By direct application of Theorem 1 and

since $P(z) = z^3+\alpha_1 z^2+\alpha_2 z+\alpha_3 = \prod_{i=1}^{3} (z-\lambda_i)$ gives $(\alpha_3, \alpha_2, \alpha_1)$ as a point in E^3

then $(z-1)(z-1)(z-1)$ gives $(-1, +3, -3)$,

$(z-1)(z-1)(z+1)$ gives $(+1, -1, -1)$,

$(z-1)(z+1)(z+1)$ gives $(-1, -1, +1)$,

and $(z+1)(z+1)(z+1)$ gives $(+1, +3, +3)$

The four points obtained are the vertices of the convex hull of $\mathscr{D}_3$ (see Fig. 3).

The convex hull of $\mathscr{D}_n$ is clearly much easier to obtain than its exact boundary.

III. STABILIZABILITY BY OUTPUT FEEDBACK-
A GEOMETRIC APPROACH

The concept of the parameter space and its application in the design of controllers is as old as the root locus method. Naimark [4] investigated stability domains for a class of continuous time systems at about the same time Evans [5] introduced the root locus method. The parameter space and stability domain methods were not developed satisfactorily and their potential as tools for investigation of important problems as the one presented in this paper has not been recognized.

In particular, the case of discrete-time linear systems received almost no serious consideration. The main contribution of this work is to reintroduce this geometric approach by using it to achieve a better understanding of the output feedback problem.

As was shown in the introduction, for single-input/single-output systems, the stabilizability question reduces to whether the straight line

$\underline{\alpha} + K\underline{\beta}$ intersects $\mathscr{D}_n$ or not. The following example illustrates the application of our geometric method.

Example 3.

Consider the unstable, but controllable and observable, system

$$A = \begin{bmatrix} 10 & 1 \\ 2 & 4 \end{bmatrix}, \; B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \; C = [1 \quad 0]$$

To establish whether or not it is stabilizable by output feedback, we transform it first to the controllable canonical form

$$\tilde{A} = \begin{bmatrix} 0 & 1 \\ -38 & 14 \end{bmatrix}, \qquad \tilde{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \text{ and } \qquad \tilde{C} = [-4 \quad 1]$$

Hence, $\underline{\alpha} = (\alpha_2, \alpha_1) = (38, -14)$ and $\underline{\beta} = (4, 1)$. As shown in Fig. 4, the locus of $\underline{\alpha} + K\underline{\beta}$ does not intersect $\mathscr{D}_2$. As a result, the given system is not stabilizable by output feedback. ☐

Next we describe briefly how to extend our investigation to single-input/multi-output and multi-input/single-output systems. The class of stabilizable systems is then characterized. This is extended to the case where robustness and tolerances are taken into consideration.

The case of single-input/multi-output systems

This case is an obvious extension of the single-input/single-output case. From the same controllable canonical form, we find $\underline{\alpha}$ and plot it in the companion parameter space with $\mathscr{D}_n$ constructed around the origin. Any selection of K determines a projection of C since K is a row in this case. Hence, at the end of the vector $\underline{\alpha}$ construct each column of C as a vector $\underline{\beta}_i$. Now K adds to $\underline{\alpha}$ a linear combination of the $\underline{\beta}_i$'s, i.e., an arbitrary vector in the hyperplane determined by the $\underline{\beta}_i$'s. The system is stabilizable by output feedback if and only if this hyperplane intersects $\mathscr{D}_n$.

The case of multi-input/single-output systems

All the above is directly applicable to this case since observability and duality allow us to transform the system into a form where A is the transpose of the form given earlier and the roles of B and C are interchanged.

Characterizing the class of stabilizable systems

A single-input/single-output discrete linear system determines uniquely a point $\underline{\alpha}$ in the companion parameter space. All systems with the same $\underline{\alpha}$, but arbitrary C are represented by the same point $\underline{\alpha}$. Figure 5 shows the cone that bounds all vectors $\underline{\beta}$ emanating from $\underline{\alpha}$ and corresponding to the class of systems stabilizable by output feedback. If the system is multi-output, then the interior of this cone must have a non-empty intersection with the hyperplane defined by the $\underline{\beta}_i$'s mentioned earlier, for the system to be stabilizable.

Incorporating robustness in the characterization of stabilizable systems

If a system is to remain stabilizable in the presence of bounded uncertainty or finite tolerance on $\underline{\alpha}$ and $\underline{\beta}$, then $\underline{\beta}$ has to exist in a tighter domain contained in the cone defined above for the system to be stabilizable by output feedback. Two examples are shown in Fig. 6. Other types of tolerance, e.g., on K, could be accounted for similarly.

IV. INTEGRAL GEOMETRY AND LIMITATIONS OF OUTPUT FEEDBACK

Given a class of linear, discrete-time systems, is it possible to measure the applicability of output feeback as a means for stabilization? This interesting question arises naturally from observing that it is "unlikely" for systems with large $\underline{\alpha}$ (in some norm) to be stabilizable by output feedback.

Consider the case of single-input/single-output systems. Fixing $\underline{\alpha}$ and selecting $\underline{\beta}$ "randomly" we can measure the probability that a random straight line passing by $\underline{\alpha}$, intersects $\mathscr{D}_n$. To give rigorous meaning to the probabilistic statements above we need the tools of geometric probability [6] or integral geometry. (Refer to [7] for a good survey and bibliography.)

Integral geometry has been applied in a number of subjects. An example is its application to pattern recognition as suggested first by Novikoff in [8], which contains a good treatment of the points that should be considered in applications of integral geometry.

We give below an example of the insight that we can get by applying geometric probability to output feedback. A rigorous treatment is left to a more detailed study in [3].

Example 4.

Consider the case of single-input/single-output second-order systems. Let $\underline{\alpha}$ be fixed and assume $\underline{\beta}$ is uniformly distributed over its angle. Now the probability of picking a system from this class at random and finding that it is stabilizable by output feedback, $P_{\underline{\alpha}}$, is easy to calculate and is inversely proportional to the angle of the cone constructed from $\underline{\alpha}$ and subtending $\mathscr{D}_2$. Figure 7 shows the upper lower and lower upper bounds of $P_{\underline{\alpha}}$ as a function of $||\underline{\alpha}||$ in the norm L_2.

It is obvious that since $\mathscr{D}_n$ is confined to a bounded region around the origin that the larger $\underline{\alpha}$ is, the smaller the percentage of systems with such $\underline{\alpha}$ that are stabilizable by output feedback and the tighter the tolerances are on this feedback and the system for it to be robust. Since $\lim_{||\underline{\alpha}||\to\infty} P_{\underline{\alpha}} = 0$, we conclude that output feedback is practically useless in stabilizing highly unstable systems, and is mainly effective if the poles of the system are not too far from the open unit disc.

V. CONCLUSION

We have developed a new geometric approach to study the stabilizability of linear systems by output feedback. One of the main advantages of this approach over the root locus method is that the stability domain $\mathscr{D}_n$ need be constructed but once and for any given system a hyperplane is constructed then we check to determine if it intersects $\mathscr{D}_n$ or not. In contrast to the root locus method where the stability domain in the complex plane is the open unit disc, every system produces a set of curves that need to be constructed. Clearly the companion parameter space is superior since it transfers the difficulty to constructing $\mathscr{D}_n$ which does not depend on the given system and is itself very easy to construct using the simple recursive technique given in Section II. The convex hull of $\mathscr{D}_n$ which is trivial to compute makes our approach even more appealing. Our results enable one to consider complete classes of systems simultaneously. This is particularly suitable for dealing with robustness. The integral geometric application in Section IV is interesting and should be extended to provide a quantitative measure of the feasibility of different control techniques.

A detailed study of the concepts introduced in this work, extended to continuous-time systems, multi-input/multi-output systems and time-varying feedback is in progress.

REFERENCES

1. B. D. O. Anderson, N. K. Bose and E. I. Jury, "Output feedback stabilization and related problems - solution via decision methods," IEEE Trans. Automat. Contr., Vol. AC-20, pp 53-66, Feb. 1975

2. H. Kimura, "Pole assignment by gain output feedback," IEEE Trans. Automat. Contr., Vol. AC-20, pp 509-516, Aug. 1975.

3. Adly T. Fam, "Theory and applications of feedback in multivariable linear systems," Ph.D. dissertation, Univ. of Calif., Irvine. (In preparation.)

4. Yu. I. Naimark, "The stability of linearized systems," (in Russian), Leningrad, 1949.

5. W. R. Evans, "Control system synthesis by root locus method," Trans. AIEE, Vol. 69, pp 1-4, Mar. 1950.

6. M. G. Kendall and P. A. P. Moran, "Geometric Probability," Charles Griffin & Co., Ltd., London, 1963.

7. G. I. Drinfeld, "Integral geometry," from "Progress in Mathematics," Vol. 12 - Algebra and geometry - edited by R. V. Gamkrelidze, Plenum Press, New York-London, 1972.

8. A. B. J. Novikoff, "Integral geometry as a tool in pattern perception," Principles of Self-Organization, H. Von Forster and G. W. Zopf, eds., Pergamon Press, New York, pp 347-368, 1962.

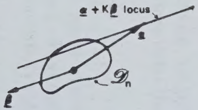

$\underline{\alpha} + K\underline{\beta}$ locus
$\underline{\alpha}$
$\underline{\beta}$
$\mathscr{D}_n$

Fig. 1. Iff the locus of $\underline{\alpha} + K\underline{\beta}$ intersects $\mathscr{D}_n$ then the system is stabilizable.

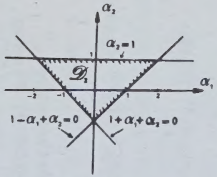

α_2
$\alpha_2 = 1$
$\mathscr{D}_2$
α_1
-2
-1
1
2
$1 - \alpha_1 + \alpha_2 = 0$
$1 + \alpha_1 + \alpha_2 = 0$

Fig. 2. Finding $\mathscr{D}_2$ from Equations (5), (6) and (7). (Example 1).

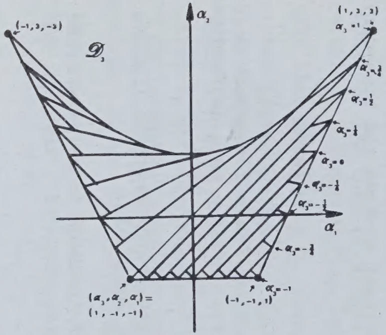

α_2
$\mathscr{D}_3$
(-1, 3, -3)
(1, 3, 3)
$\alpha_3 = 1$
$\alpha_3 = \frac{3}{4}$
$\alpha_3 = \frac{1}{2}$
$\alpha_3 = \frac{1}{4}$
$\alpha_3 = 0$
$\alpha_3 = -\frac{1}{4}$
$\alpha_3 = -\frac{1}{2}$
α_1
$\alpha_3 = -\frac{3}{4}$
$\alpha_3 = -1$
$(\alpha_3, \alpha_2, \alpha_1) =$
(1, -1, -1)
(-1, -1, 1)

Fig. 3. $\mathscr{D}_3$ and vertices of its convex hull. The triangular contours are produced from $\mathscr{D}_1$ and $\mathscr{D}_2$. (Example 2)

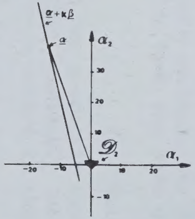


Fig. 4. An unstabilizable system. (Example 3)

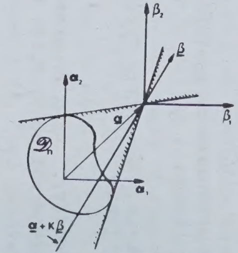

β_2
$\underline{\beta}$
α_2
β_1
$\underline{\alpha}$
$\mathscr{D}_n$
α_1
$\underline{\alpha} + K\underline{\beta}$

Fig. 5. Cone of stabilizable systems.

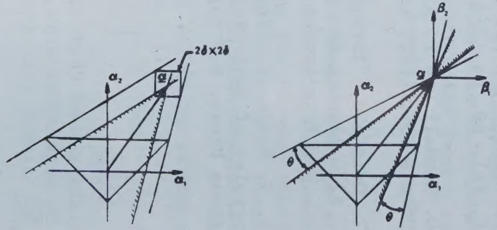

$2\delta \times 2\delta$
α_2
$\underline{\alpha}$
α_1
β_2
$\underline{\alpha}$
α_2
β_1
θ
α_1
θ

Fig. 6. Effect of tolerances of $\pm\delta$ on $\underline{\alpha}$ and $\pm\theta$ on the angle of $\underline{\beta}$, on the angle of the cone of stabilizable systems. Shown for second order systems.

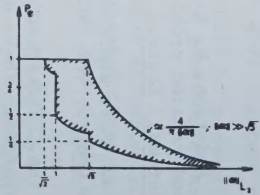


Fig. 7. Bounds of probability of finding a system stabilizable. (Example 4)

AUTOMATIC GENERATION CONTROL: A DECENTRALIZED APPROACH

M. S. Ćalović
Serbian Electric Power Board
Belgrade, Yugoslavia

M. V. Đorović
City of Belgrade Headquarters
Belgrade, Yugoslavia

D. D. Šiljak
University of Santa Clara
Santa Clara, California

ABSTRACT

A decentralized scheme is proposed for a piece-by-piece design of Automatic Generation Control (AGC) of interconnected power systems. The scheme can take advantage of the special structural properties of power systems and come up with a considerable reduction of dimensionality problems in AGC regulator design.

1. INTRODUCTION

The purpose of this paper is to present a new method for decentralized design of Automatic Generation Control (AGC) for interconnected power systems. The method is based upon the decomposition-aggregation theory of large-scale dynamic systems [1, 2] and the modified linear regulators [3, 4] designed for individual power areas. A strong motivation for developing such a method comes from two important facts. First, the same "one control per area" principle is used in the conventional AGC design, and we are able to provide a rigorous basis for such practice. Second, in the modern multivariable AGC schemes with centralized design and/or operation (see the survey in [5]), the whole interconnected model is manipulated "in one piece" in order to obtain a global control for the overall system. This creates both conceptual and computational difficulties whenever a control of large power systems is attempted mainly because centralized information has to be obtained from power areas and generating plants which are spread over large geographic territories and, at the same time, processed in a "one shot" design procedure. In the present decentralized scheme, the AGC regulators are designed "piece-by-piece" utilizing the individual area models and applying the concept of autonomous local control to each power area and the associated tie-lines. Then, the compatibility of the local AGC regulators on the overall system level is established by the decomposition-aggregation method for control of large composite systems.

A conceptually interesting novelty in the decentralized AGC scheme, is a new notion of the subsystem. A power system is decomposed so that each subsystem represents an area and all the tie-lines originating from that area. Therefore, each subsystem consists of all the states accessible from the corresponding area, and a certain "overlapping" exists among the subsystems which was not present in the original decomposition-aggregation concept [1, 2]. Such decomposition scheme leads to a decentralized area control of proportional-plus-integral type which is designed as a linear-quadratic regulator with complete subsystem feedback. It retains the area control concept and load distribution property of the conventional regulators while, at the same time, guarantees improved system transients and stability margins for the overall interconnected power system. In the paper, we discuss various properties of decentralized AGC regulators, as well as the problems of their implementation.

Now, we summarize the main features of the proposed decentralized AGC scheme:

(i) Each decentralized regulator applies only the local feedback based upon the measurements accessible from its own area (this should include only measurements usually available in control centers);

(ii) In the local regulator design, only models of the corresponding area model and the associated tie-lines are used;

(iii) The local control law is a linear, constant feedback, independent of the area load disturbance;

(iv) The steady-state errors of the frequency and tie-line exchange are reduced to zero;

(v) In comparison with actually used conventional AGC, the decentralized AGC regulators produce improved system transients and increased stability margins;

(vi) A desired steady-state distribution of the generated power is incorporated in the decentralized AGC scheme as an independent process. This implies the existence of only one independent local control per area.

2. A DECENTRALIZED MODEL

A linearized model of an interconnected power system can be described by the equations

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) + Fz(t) \\ y(t) &= C^T x(t) \quad , \end{aligned} \tag{1}$$

where x, u, z, and y are n, m, ℓ, and r vectors of states, inputs, unmeasurable disturbances, and outputs, respectively, A, B, F, and C are constant matrices of appropriate dimensions, and $x(t_0) = 0$. The model (1) is composed of a set of ℓ area models and the associated tie-line models, which can be presented in a decomposed form as

$$\begin{aligned} \dot{x}_i(t) &= A_i x_i(t) + a_{ti}\Delta p_{ei}(t) + B_i u_i(t) + f_i z_i(t) \\ \Delta\dot{p}_{ei}(t) &= \sum_{\substack{j=1 \\ j\neq i}}^{\ell} [m_{ij}^T x_i(t) - m_{ji}^T x_j(t)] \\ y_i(t) &= C_i^T x_i(t) \quad , \qquad\qquad i = 1, 2, \dots , \ell \; . \end{aligned} \tag{2}$$

In (2), x_i, u_i, z_i, and y_i are n_i, m_i, 1, and r_i vectors of the i-th area states, inputs, disturbances, and outputs, respectively, and Δp_{ei} is the variation of the total power exchange of the i-th area, such that

$$\Delta p_{ei} = \sum_{\substack{j=1 \\ j\neq i}}^{\ell} \Delta p_{ij} \; , \quad i = 1, 2, \dots , \ell \tag{3}$$

where Δp_{ij} are power exchange deviations between the i-th and the j-th area expressed in per unit. The matrices A_i, B_i, f_i, and C_i are constant matrices of appropriate dimensions, a_{ti} and m_{ij} are constant coupling n_i vectors, and m_{ji} is a constant coupling n_j vector. Finally, in addition to $x_i(t_0) = 0$, we assume $\Delta p_{ei}(t_0) = 0$.

The particular properties of the model represented by equations (1)-(3) imply the following relations:

$$n = \sum_{i=1}^{\ell} n_i + \ell - 1 \; , \quad m = \sum_{i=1}^{\ell} m_i \; , \quad r = \sum_{i=1}^{\ell} r_i \tag{4}$$

and

$$\sum_{j=1}^{\ell} \Delta p_{ej}/\alpha_{ij} = 0 \ , \tag{5}$$

where $\alpha_{ij} = P_{i0}/P_{j0}$ is the size-ratio coefficient for the j-th area relative to the steady-state load of the i-th area taken as reference.

To design the decentralized area AGC regulators, the decoupled model (2) should be suitably rearranged. First, in order to achieve the design objectives (iii) and (iv), it is necessary to augment the model (2) with the new state defined as an integral of the dynamic Area Control Error (ACE) y_i^o as proposed in [3, 4]. We have for all $i = 1, 2, \ldots , \ell$,

$$v_i(t) = v_{i0} + \int_{t_0}^{t} y_i^o(t)dt \ , \tag{6}$$

where $y_i^o = \Delta p_{ei} + b_{si}\Delta f_i$, b_{si} is the bias factor and Δf_i is the frequency deviation in the i-th area (expressed in per unit). The expression for y_i^o can be rewritten as $y_i^o = \Delta p_{ei} + d_i^T x_i$, where d_i is a constant column n vector.

The second rearrangement is a consequence of the design objective (vi), which is known as the Area Control Concept. A desired (unique) area control w_i is obtained by introducing the equality constraints to the unit control vectors u_i in (2) of the type

$$u_i(t) = \nu_i w_i(t) \ , \tag{7}$$

where ν_i is the constant participation m_i vector [3, 4], the elements of which satisfy the relation

$$\sum_{k=1}^{m_i} \nu_k = 1 \ , \tag{8}$$

so that it determines the relative steady-state load distribution in the i-th area.

With regard to equations (6) and (7), the interconnected power system model (2) can be rewritten as

$$\begin{aligned} \dot{x}_i(t) &= A_i x_i(t) + a_{ti}\Delta p_{ei}(t) + B_i \nu_i w_i + f_i z_i(t) \\ \dot{v}_i(t) &= d_i^T x_i(t) + \Delta p_{ei}(t) \\ \Delta \dot{p}_{ei}(t) &= \sum_{\substack{j=1 \\ j\neq i}}^{\ell} [m_{ij}^T x_i(t) - m_{ji}^T x_i(t)] \\ & \qquad\qquad i = 1, 2, \ldots , \ell \end{aligned} \tag{9}$$

where $x_i(t_0) = 0$, $v_i(t_0) = v_{i0}$, and $\Delta p_{ei}(t_0) = 0$.

The equations (9) define an $(n_i + 2)$ dimensional model of one subsystem in the sense of the decentralized system representation. Assuming that $z(t)$ is constant, the model (9) can be reduced to the standard form [1, 2] of a decomposed decentralized system

$$\dot{\hat{x}}_i(t) = \hat{A}_i \hat{x}_i(t) + \hat{b}_i \hat{w}_i + \sum_{\substack{j=1 \\ j\neq i}}^{\ell} \hat{A}_{ij}\hat{x}_j(t) \ , \qquad i = 1, 2, \ldots , \ell \tag{10}$$

where

$$\hat{x}_i = (x_i - x_{iss}, v_i - v_{iss}, \Delta p_{ei})^T, \ \hat{w}_i = w_i - w_{iss} \tag{11}$$

and

$$\hat{A}_i = \left[\begin{array}{c|c|c} A_i & 0 & a_{ti} \\ \hline d^T & 0 & 1 \\ \hline \sum\limits_{j\neq i}^{\ell} m_{ij}^T & 0 & 0 \end{array}\right] , \quad \hat{b}_{ij} = \left[\begin{array}{c} B_i\nu_i \\ \hline 0 \\ \hline 0 \end{array}\right] , \quad \hat{A}_{ij} = \left[\begin{array}{c|c|c} 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline -m_{ji}^T & 0 & 0 \end{array}\right] \qquad (12)$$

with $\hat{x}_{i0} = (-x_{iss}, v_{i0} - v_{iss}, 0)^T$, where the subscript "ss" designates a steady-state value of the indicated variable.

From (12), we see that there is an "overlapping" of the subsystem state vector x_i due to the fact that each subsystem in (12) incorporates the corresponding tie-line models. Therefore, the subsystem state vector contains all accessible states from the individual area it represents. Such definition of a subsystem is quite natural. Subsystems without overlapping would lead to a complete separation of an interconnected power system, which would aggrevate the decentralized AGC regulator design.

3. REGULATOR DESIGN

Following the design objectives (i)-(iii) in Section 1, we require that the AGC regulator should be of a decentralized type with linear constant feedback. Using the complete state feedback applied to the decomposed system model (10), we choose the local control for the i-th area as

$$\hat{w}_i(t) = -\hat{k}_i^T\hat{x}_i(t) \ , \qquad (13)$$

where

$$\hat{k}_i = (k_{pi}^T, k_{Ii}, k_{Ti})^T \qquad (14)$$

is an $(n_i + 2)$-vector of feedback gains associated with the i-th area. In (14), we have:

k_{pi} - n_i vector of proportional control gains;

k_{Ii} - integral control gain;

k_{Ti} - proportional control gain for the net interchange load deviation.

By substituting the local control (13) into (10), we get the closed-loop system as

$$\dot{\hat{x}}_i(t) = (\hat{A}_i - \hat{b}_i\hat{k}_i^T)\hat{x}_i(t) + \sum_{\substack{j=1 \\ j\neq i}}^{\ell} \hat{A}_{ij}\hat{x}_j(t) \qquad i = 1, 2, \dots , \ell \ . \qquad (15)$$

A choice of the gains vectors $\hat{k}_i$ which produces a closed-loop system (15) with a satisfactory degree of exponential stability, can be made using the stabilization procedure of reference [2]. Since results are already available for power system load and frequency control of individual area by utilizing the optimum linear regulator with integral feedback [3], we choose $\hat{k}_i$'s in (15) to optimize locally each subsystem without the interconnections (set $\sum_{j=1}^{\ell} A_{ij}x_j \equiv 0$) . With this choice of gains $\hat{k}_i$, we establish stability of the overall closed-loop system (15) applying the concept of vector Liapunov function as proposed in [2].

With each decoupled subsystem

$$\dot{\hat{x}}_i(t) = \hat{A}_i\hat{x}_i(t) + \hat{b}_i\hat{w}_i(t) \ , \quad i = 1, 2, \dots , \ell \qquad (16)$$

we associate a quadratic performance index

$$\hat{J}_i(t_0, \hat{x}_{i0}, \hat{w}_i) = \int_{t_0}^{\infty} e^{2\pi t}(\hat{x}_i^T\hat{Q}_i\hat{x}_i + \hat{r}_i\hat{w}_i^T\hat{w}_i)dt \ , \tag{17}$$

where $\hat{Q}_i$ is a symmetric nonnegative definite matrix $(\hat{Q}_i = \hat{Q}_i^T \geq 0)$ and $\hat{r}_i$ is a positive number $(\hat{r}_i > 0)$. We choose the matrix $\hat{Q}_i$ as a quasi-diagonal matrix,

$$\hat{Q}_i = \left[\begin{array}{c:c:c} Q_{Pi} & 0 & 0 \\ \hdashline 0 & q_{Ii} & 0 \\ \hdashline 0 & 0 & q_{Ti} \end{array}\right] \tag{18}$$

where $Q_{Pi} = Q_{Pi}^T \geq 0$ is a constant $n_i \times n_i$ matrix and $q_{Pi} > 0$, $q_{Ti} \geq 0$ are scalars. In (17), π is a positive number which is a measure of the degree of exponential stability.

It is a well-known fact [6] that under the assumption that the pair $(\hat{A}_i, \hat{b}_i)$ is completely controllable, there is a unique optimal gain vector $\hat{k}_i^*$ for the feedback control law in (13), given as

$$\hat{k}_i^{*T} = \hat{r}_i^{-1}\hat{b}_i^T\hat{P}_i \ , \tag{19}$$

where $\hat{P}_i$ is an $(n_i+2)\times(n_i+2)$ symmetric positive definite matrix which is the solution of the Riccati equation

$$\hat{P}_i(\hat{A}_i + \pi I_i) + (\hat{A}_i + \pi I_i)^T \hat{P}_i - \hat{r}_i^{-1}\hat{P}_i\hat{b}_i\hat{b}_i^T\hat{P}_i + \hat{Q}_i = 0 \ , \tag{20}$$

where I_i is the $(n_i+2)\times(n_i+2)$ identity matrix. The optimal control for the i-th subsystem is

$$\hat{w}_i^*(t) = \hat{k}_i^{*T}\hat{x}_i(t) \ , \tag{21}$$

which results in the optimal cost

$$\hat{J}_i^*(t_0, \hat{x}_{i0}) = e^{-2\pi t_0}\,\hat{x}_{i0}^T\hat{P}_i\hat{x}_{i0} \ . \tag{22}$$

It is equally well-known fact [6] that under the assumption that $\hat{Q}_i$ can be factored as $\hat{Q}_i = \hat{M}_i\hat{M}_i^T$, where $\hat{M}_i$ is an $(n_i+2)\times(n_i+2)$ matrix, so that the pair $(A_i, \hat{M}_i)$ is completely observable, each closed-loop subsystem without interactions,

$$\dot{\hat{x}}_i(t) = (\hat{A}_i - \hat{r}_i^{-1}\hat{b}_i\hat{b}_i^T\hat{P}_i)\hat{x}_i(t) \ , \quad i = 1, 2, \ldots, \ell \tag{23}$$

is globally exponentially stable. That is, there exists a positive number Π_i such that the solution $\hat{x}_i(t;\, t_0, \hat{x}_{i0})$ satisfies the inequality

$$||\hat{x}_i(t;\, t_0, \hat{x}_{i0})|| \leq \Pi_i||\hat{x}_{i0}||\,\exp[-\pi(t-t_0)] \ , \tag{24}$$

for all t_0 and $\hat{x}_{i0}$. Now, it remains to limit properly the interconnections $\hat{A}_{ij}\hat{x}_j(t)$ in (15), and from exponential stability of each subsystem (23), imply the same stability property of the overall system (15). Before we set out to do this in the next section, let us identify the individual components in (14) for the optimal gain vector $\hat{k}_i^*$ given by (19).

First, we partition the matrix $\hat{P}_i$ as

$$\hat{P}_i = \left[\begin{array}{c:c:c} p_{i11} & p_{i12} & p_{i13} \\ \hdashline p_{i12}^T & p_{i22} & p_{i23} \\ \hdashline p_{i13}^T & p_{i23} & p_{i33} \end{array}\right] . \qquad (25)$$

Then, the components of $\hat{k}_i^*$ are obtained from (19) and (25) as

$$k_{Pi}^{*T} = \hat{r}_i^{-1} \nu_i^T B_i^T P_{i11}, \; k_{Ii}^* = \hat{r}_i^{-1} \nu_i^T B_i^T p_{i12}, \; k_{Ti}^* = \hat{r}_i^{-1} \nu_i^T B_i^T p_{i13} . \qquad (26)$$

The optimal control law (21) reduces to the standard Proportional-Plus-Integral autonomous area control

$$\hat{w}_i^*(t) = k_{Pi}^{*T} x_i(t) - k_{Ii}^* \int_{t_0}^{t} [d_i^T x_i(t) + \Delta p_{ei}(t)]dt - k_{Ti}^* \Delta p_{ei}(t) + w_{i0}^*, \qquad (27)$$

where the initial control is $w_{i0}^* = \hat{r}_i^{-1} \nu_i^T B_i^T P_{i12} \upsilon_{i0}$.

A block diagram representation of a decentralized AGC regulator for the i-th area is shown in the figure.

4. OVERALL SYSTEM STABILITY

With the partition (14) and the components (26) of the optimal gain vector of (19), the overall system (15) can be rewritten as

$$\dot{x}_i(t) = (\hat{A}_i - \hat{b}_i \hat{k}_i^{*T}) x_i(t) + \sum_{j=1}^{\ell} \hat{A}_{ij} \hat{x}_j(t), \; i = 1, 2, \dots , \ell \qquad (28)$$

where

$$\hat{A}_i - \hat{b}_i \hat{k}_i^{*T} = \left[\begin{array}{c:c:c} \hat{A}_i - B_i \nu_i k_{Pi}^{*T} & -B_i \nu_i k_{Ii}^* & a_{ti} - B_i \nu_i k_{Ti}^* \\ \hdashline d_i^T & 0 & 0 \\ \hdashline \sum\limits_{j \neq i}^{\ell} m_{ij}^T & 0 & 0 \end{array}\right], \; \hat{A}_{ij} = \left[\begin{array}{c} 0 \\ \hdashline 0 \\ \hdashline -m_{ji}^T \end{array}\right] . \qquad (29)$$

In order to establish stability of the closed-loop system (28), we use the vector Liapunov function concept [1, 2]. With each optimal free subsystem

$$\dot{\hat{x}}_i(t) = (\hat{A}_i - \hat{r}_i^{-1} \hat{b}_i \hat{b}_i^T \hat{P}_i) \hat{x}_i(t) , \quad i = 1, 2, \dots , \ell \qquad (23)$$

we associate a scalar Liapunov function

$$v_i(\hat{x}_i) = (\hat{x}_i^T \hat{P}_i \hat{x}_i)^{\frac{1}{2}} \qquad (30)$$

where the symmetric positive definite matrix $\hat{P}_i$ is computed as a solution of equation (20). Using equation (20), it is easy to show that the function $v_i(\hat{x}_i)$ has the following estimates:

$$\eta_{i1} ||\hat{x}_i|| \leq v_i(\hat{x}_i) \leq \eta_{i2} ||\hat{x}_i||$$

$$\dot{v}_i(\hat{x}_i)_{(23)} \leq -\eta_{i3} v_i(\hat{x}_i)$$

$$||\text{grad } v_i(\hat{x}_i)|| \leq \eta_{i4} \quad , \qquad (31)$$

where

$$\eta_{i1} = \lambda_m^{\frac{1}{2}}(\hat{P}_i) \, , \qquad \eta_{i2} = \lambda_M^{\frac{1}{2}}(\hat{P}_i)$$

$$\eta_{i3} = \pi + \frac{1}{2}\, \lambda_m(\hat{S}_i)\lambda_M^{-1}(\hat{P}_i) \, , \quad \eta_{i4} = \lambda_m^{-\frac{1}{2}}(\hat{P}_i)\lambda_M(\hat{P}_i) \tag{32}$$

and

$$\hat{S}_i = \hat{r}_i^{-1}\hat{P}_i\hat{b}_i\hat{b}_i^T\hat{P}_i + \hat{Q}_i \quad . \tag{33}$$

In (31), $\dot{v}_i(\hat{x}_i)_{(23)}$ denotes the time derivative of the function $v_i(\hat{x}_i)$ calculated with respect to the corresponding equation (23). In (32), λ_m and λ_M are the minimum and maximum eigenvalues of the indicated matrices. By integrating the second inequality and using the first inequality in (31) we can get inequality (24) with $\Pi_i = \eta_{i1}^{-1}\eta_{i2}$. Therefore, estimates (31) confirms the fact that each subsystem (23) is exponentially stable with the degree π .

Now, we compute the time derivative $\dot{v}_i(\hat{x}_i)_{(28)}$ of the function $v_i(\hat{x}_i)$ with respect to the interconnected subsystems (28). We make use of (32), (33), and the inequality

$$|| \sum_{j=1}^{\ell} \hat{A}_{ij}\hat{x}_j || \leq \sum_{j=1}^{\ell} \xi_{ij}||\hat{x}_j|| \, , \tag{34}$$

where $\xi_{ij} = \lambda_M^{\frac{1}{2}}(\hat{A}_{ij}^T\hat{A}_{ij})$, to get

$$\begin{aligned} \dot{v}_i(\hat{x}_i)_{(28)} &= -\dot{v}_i(\hat{x}_i)_{(23)} + [\text{grad } v_i(\hat{x}_i)]^T \sum_{\substack{j=1 \\ j\neq i}}^{\ell} \hat{A}_{ij}\hat{x}_j \\ &\leq -\eta_{i3}v_i(\hat{x}_i) + \eta_{i4} \sum_{\substack{j=1 \\ j\neq i}}^{\ell} \xi_{ij}\eta_{j1}^{-1}v_j(\hat{x}_j), \quad i = 1, 2, \ldots , \ell \, . \end{aligned} \tag{35}$$

By introducing the vector Liapunov function $v(\hat{x})$ defined by

$$v = (v_1, v_2, \ldots , v_\ell) \, , \tag{36}$$

and $\hat{x} = (\hat{x}_1^T, \hat{x}_2^T, \ldots , \hat{x}_\ell^T)^T$, we can rewrite scalar inequalities in (35) as the vector differential inequality

$$\dot{v} \leq Wv \, , \tag{37}$$

where the $\ell \times \ell$ constant matrix $W = (w_{ij})$ is defined as

$$w_{ij} = \begin{cases} -\eta_{i3} \, , & i = j \\ \xi_{ij}\eta_{i4}\eta_{j1}^{-1} \, , & i \neq j \end{cases} \quad . \tag{38}$$

The inequality (37) represents an aggregate model for the overall system (28), and it is a well-known fact [1, 2] that stability of the aggregate matrix W (that is, all eigenvalues of W have negative real parts) implies global asymptotic stability of the equilibrium $x = 0$ of the interconnected system (28). Since the matrix W has nonnegative off-diagonal elements $(a_{ij} \geq 0, \; i \neq j)$, it is stable if and only if the Sevastyanov-Kotelyanskii conditions

$$(-1)^k \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1k} \\ w_{21} & w_{22} & \cdots & w_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ w_{k1} & w_{k2} & \cdots & w_{kk} \end{bmatrix} > 0 \, , \quad k = 1, 2, \ldots , \ell \tag{39}$$

are satisfied [1]. Using reference [2] we can also show that conditions (39) imply exponential stability of the overall system (28) and, furthermore, compute an estimate $\hat{\pi}$ of the degree of exponential stability of the system (28). From [1], we recall that for a matrix W with nonnegative diagonal elements, conditions (39) are equivalent to saying that there exist positive numbers d_i, $i = 1, 2, \ldots, \ell$, and $\hat{\pi}$ such that

$$|w_{jj}| - d_j^{-1} \sum_{\substack{i=1 \\ i \neq j}}^{\ell} d_i |W_{ij}| \geq \hat{\pi}, \quad j = 1, 2, \ldots, \ell \tag{40}$$

that is, W is a quasidominant diagonal matrix. Using (37) and (40) we can consider the function

$$\nu(\hat{x}) = \sum_{i=1}^{\ell} d_i v_i(\hat{x}_i), \tag{41}$$

as a single scalar Liapunov function for the system (28), and compute the derivative of $\nu(\hat{x})$ with respect to (28) as

$$\begin{aligned} \dot{\nu}(\hat{x})_{(28)} &\leq \sum_{i=1}^{\ell} d_i \sum_{j=1}^{\ell} w_{ij} v_j(\hat{x}_j) \\ &\leq \sum_{j=1}^{\ell} v_j(\hat{x}_j) \sum_{i=1}^{\ell} d_i w_{ij} \\ &\leq -\sum_{j=1}^{\ell} d_j |w_{jj}| v_j(\hat{x}_j) + \sum_{j=1}^{\ell} v_j(\hat{x}_j) \sum_{\substack{i=1 \\ i \neq j}}^{\ell} d_i |w_{ij}| \\ &\leq -\pi \nu(\hat{x}) . \end{aligned} \tag{42}$$

By integrating the last inequality in (42), we get

$$\nu(\hat{x}) \leq \nu(\hat{x}_0) \exp[-\pi(t - t_0)] . \tag{43}$$

Finally, by using the well-known inequalities involving norms and quadratic forms [1], inequality (43) can be rewritten in terms of the solution $\hat{x}(t; t_0, \hat{x}_0)$ of the overall system (28) as

$$||\hat{x}(t; t_0, \hat{x}_0)|| \leq \hat{\Pi} ||\hat{x}_0|| \exp[-\hat{\pi}(t - t_0)] \tag{44}$$

where $\hat{\Pi} = \ell^{\frac{1}{2}} d_m^{-1} d_M \eta_{m1}^{-1} \eta_{M2}$, $d_m = \min_i d_i$, $d_M = \max_i d_i$, $\eta_{m1} = \min_i \eta_{i1}$, $\eta_{M2} = \max_i \eta_{i2}$. Inequality (44) is valid for all t_0, x_0, and t and, thus implies global exponential stability of the closed-loop system (28).

5. CONCLUSIONS

A decentralized scheme was developed to reduce the complexity in the design of Automatic Generation Control for interconnected power systems. Besides useful numerical aspects, the scheme is attractive also because it preserves autonomy of each interconnected area by requiring only measurements and use of local control variables.

In the present decentralized scheme, a complete-state feedback is applied to control each interconnected area. This may be an overly restrictive requirement. Further research is needed to alleviate this problem and replace the state feedback by simpler output feedback schemes.

There is a subtle interplay between the choice of the performance index for the subsystems and the nature of their interactions. A suitable balance between the two aspects should be achieved to come up with a successful design. Furthermore, we know that the systems stabilized by the decomposition-aggregation method are connectively stable. That is, the subsystems can be disconnected from each other and connected together in various ways during operation without affecting stability of the overall system. At this time we do not know how to interpret connective stability in the context of the interconnected power model used in this paper due to "overlapping" of the subsystems. Additional effort is needed to explore the relationship between the subsystem properties and their interactions and obtain new schemes for decentralized design of Automatic Generation Control of large-scale power systems.

REFERENCES

[1] Šiljak, D. D., "Stability of Large-Scale Systems Under Structural Perturbations", IEEE Transactions, SMC-2(1972), 657-663.

[2] Šiljak, D. D., and M. B. Vukčević, "Large-Scale Systems: Stability, Complexity, Reliability", Journal of the Franklin Institute, 301(1976), 49-69.

[3] Ćalović, M. S., "Power Systems Load and Frequency Control Using an Optimum Linear Regulator with Integral Feedback", Fifth IFAC Congress, Paper No. 7.3, Paris, June 12-17, 1972.

[4] Ćalović, M. S., "Linear Regulator Design for a Load and Frequency Control", IEEE Transactions, PAS-91(1972), 2271-2285.

[5] Reddoch, T. W., "Automatic Generation Control/Normal: Load-Frequency Control Performance Criteria with Reference to the Use of Advanced Control Theory", Proceedings of the Engineering Foundation Conference on Systems Engineering for Power: Status and Prospects, Henniker, New Hampshire, 1975, pp. 15-26.

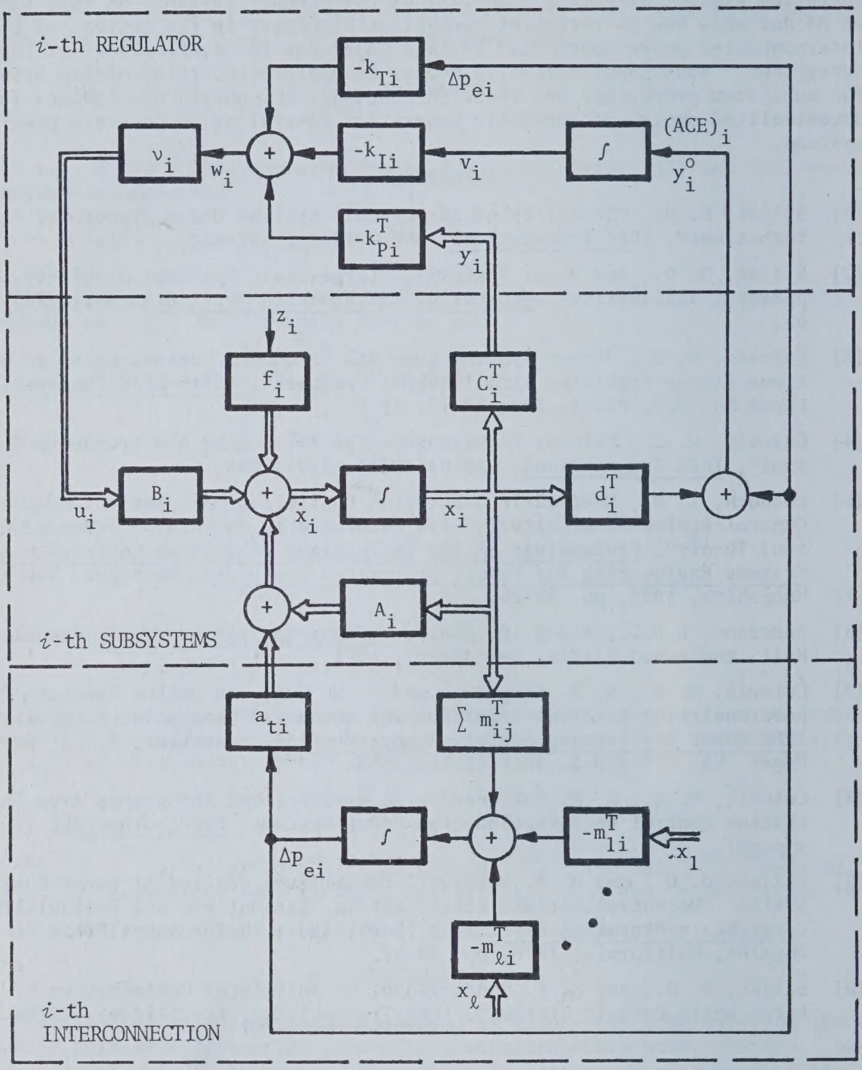
[6] Anderson, B.D.O., and J. B. Moore, "Linear Optimal Control", Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[7] Ćalović, M. S., S. P. Bingulac, and N. M. Ćuk, "An Output Feedback Proportional-Plus-Integral Regulator for Automatic Generation Control", IEEE Power Engineering Society Summer Meeting, Vancouver, B.C., Canada, Paper No. C73-489-2, July 15-19, 1973.

[8] Ćalović, M. S., N. M. Ćuk, and M. V. Đorović, "An Autonomous Area Generation Control of Interconnected Power Systems, Proceedings IEE (to appear).

[9] Šiljak, D. D., and M. B. Vukčević, "Multilevel Control of Large-Scale Systems: Decentralization, Stabilization, Estimation, and Reliability", Large-Scale Dynamical Systems, R. Saeks (ed.), Point Lobos Press, Los Angeles, California, 1976, pp. 34-57.

[10] Šiljak, D. D., and M. K. Sundareshan, "A Multilevel Optimization of Large-Scale Dynamic Systems", IEEE Transactions, AC-21(1976), 79-84.


i-th REGULATOR
$-k_{Ti}$
Δp_{ei}
ν_i
w_i
+
$-k_{Ii}$
v_i
$\int$
$(ACE)_i$
y_i^o
$-k_{Pi}^T$
y_i
z_i
f_i
C_i^T
B_i
u_i
+
$\dot{x}_i$
$\int$
x_i
d_i^T
+
A_i
+
i-th SUBSYSTEMS
a_{ti}
$\sum m_{ij}^T$
Δp_{ei}
$\int$
+
$-m_{1i}^T$
x_1
$-m_{\ell i}^T$
x_ℓ
i-th INTERCONNECTION

THE i-th AGC REGULATOR

AN APPROACH TO POLE ASSIGNMENT BY EXTERIOR ALGEBRA

V. SESHADRI and M. SAIN
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana 46556

ABSTRACT

The utility of the matrix fraction description ND^{-1} for time-invariant multivariable linear systems can be attributed in great measure to the properties of the determinant $|D|$ of the operator D. Consequently, exterior algebras designed explicitly for the analysis of such objects can be of assistance in the system theory associated with matrix fractions. In this paper, we present an initial result on the application of such algebras to the question of spectrum assignment in the minimal design problem for linear multivariable systems.

INTRODUCTION

The design of compensators in multivariable control problems often leads to the minimal design problem (MDP), which can be represented by the equation

$$G_1(v)G(v) = G_2(v). \tag{1}$$

Here, G_1, G_2 and G are all matrices over the field $F(v)$ of rational functions in the indeterminate v, with coefficients from the field F. G_1 and G_2 are known, while G represents the compensator to be designed. The minimal design problem is to enumerate all realizable G's which satisfy (1), and to pick from among them a G having realization with the least number of dynamical elements.

A transfer function matrix $G(v)$ is "usually" regarded as a numerator polynomial matrix and a denominator polynomial, that is, as $N(v)/d(v)$. The problem with such a denominator polynomial is that it does not tell us about the characteristic polynomial in a minimal realiztion of G, but rather about the minimal polynomial. One side effect of this is that the "usual" view of transfer function matrices does not directly tell us how many dynamical elements are required to realize it.

An alternative to the "usual" representation is the use of a numerator polynomial matrix and a denominator polynomial matrix. Thus, we represent the matrix G in terms of matrix fractions, as $N(v)D^{-1}(v)$, where N and D, the numerator matrix and denominator matrix, respectively, are matrices over the principal ideal domain ring $F[v]$ of polynomials in v with coefficients in $F[1]$. The above representation, by itself, is rather trivial; the interesting thing to do is to remove common factors from N and D. Suppose N and D have the representations $\hat{N}\underline{R}$ and $\hat{D}\underline{R}$, respectively, where $\underline{R}$ is a polynomial matrix. Then, we can remove $\underline{R}$ from both the numerator and denominator because

$$ND^{-1} = [\hat{N}\underline{R}][\hat{D}\underline{R}]^{-1} = \hat{N}\hat{D}^{-1},$$

a valid calculation over $F(v)$. It is conceptually easy to remove the "biggest" $\underline{R}$ possible. For a proof and description of the method for doing this, we refer the reader to [2]. We focus attention, instead, on the fact that, with a matrix fraction representation ND^{-1} of the transfer function matrix, $|D|$ is nothing but the characteristic polynomial in a minimal realization multiplied, at most, by a nonzero element of the field F, provided the only common factors $\underline{R}$ between N and D are unimodular. For a more detailed treatment of the matrix fraction idea, and to compare different approaches to the problem, the reader is referred to Wang and Davison [3], Rosenbrock [4], Popov [5], Wolovich [6] and Forney [7].

MDP: A FREE-MODULAR VIEW

Building on the work of Wang and Davison [3] and Forney [7], Sain [8] showed how MDP can be understood and solved as a problem of finding the kernel of a morphism of free F[v]-modules.

A variety of attacks can be used to convert MDP to the free-modular structure. For example, let

$$G_i(v) = \frac{N_i(v)}{d_i(v)}; \qquad i = 1, 2$$

where $d_i(v) \in F[v]$, $i = 1, 2$, and N_1 and N_2 are matrices over the same ring. Using the matrix fraction representation ND^{-1} for G, we convert (1) directly to

$$[d_2(v)N_1 \vdots -d_1(v)N_2] \begin{bmatrix} N \\ D \end{bmatrix} = [0], \tag{2}$$

whereby we see that pairs of corresponding columns in the matrices N and D are actually elements in the kernel of the morphism whose matrix is

$$T = [d_2N_1 \vdots -d_1N_2]. \tag{3}$$

In this context, Sain [8] has shown how to (i) solve for the kernel, (ii) search the kernel for elements which can be used to construct realizable G's, and (iii) among the solutions of (ii), select kernel elements whose minimal realization requires an absolute minimum number of dynamical elements. Further, software has been constructed to study the application of these ideas to numerical problems [9].

MOTIVATION FOR USE OF EXTERIOR ALGEBRA

In this paper, we consider the imposition of additional specifications that are related to the position of poles in the solutions to MDP, namely (iv) determine any fixed poles in the solutions, and (v) develop algorithms for pole assignment. We start, for this purpose, with elements in the kernel of (3) which can be used to construct realizable G's, that is, with the solutions of (ii) above.

In order that the reader may get a feel for the type of elements with which we deal, we detail, in the following, the process of solving (i) and (ii) above which, incidentally, constitute Steps 1 and 2 of Sain's MDP Algorithm [8]. Solving for the kernel is simple enough; we merely append the identity matrix to the operator matrix (3) by extending the columns, and use elementary transformations to convert the resultant matrix into an echelon form. Thus, we obtain a basis

$$b_i', \qquad 1 \le i \le s = r(\ker T)$$

where r(Ker T) denotes the rank of Ker T. Since we are interested only in those elements of the kernel which can be used to construct realizable G's, the basis set must be compatible with our desire that the elements admit a linear dynamical interpretation; the set b_i' will not, in general, satisfy this requirement, and we must apply a reduction algorithm to the basis elements in order to form a reduced basis

$$b_i, \quad 1 \leq i \leq s. \tag{4}$$

Suppose we desire a particular characteristic polynomial such as

$$|D| = |\tilde{d}_1 \tilde{d}_2 \ldots \tilde{d}_n|. \tag{5}$$

The columns of D must be linear combinations of the corresponding lower portions of the basis (4). Also, the determinant is an alternating multilinear function. Hence, it would be worthwhile to address issues (iv) and (v) in the context of an exterior algebra designed explicitly for the analysis of objects such as determinants [1].

A RELEVANT EXTERIOR ALGEBRA

A necessity for brevity precludes any lengthy introduction to the exterior algebra here. We can, however, make some brief observations while establishing a notation.

To form an exterior algebra useful in the pole assignment problems, we begin with a graded module M, or just a sequence M_i, $i = 0, 1, 2\ldots$, of $F[v]$-modules. An element of M must be an element of one of the M_i. For our purposes, M_o is taken to be $F[v]$, M_1 is taken to be a free $F[v]$-module of rank m, and

$$M_{i+1} = M_1 \otimes M_i, \quad i = 1, 2, \ldots .$$

Equipped with the tensor product for its elements, M is called the graded (tensor) algebra of M_1. The graded tensor algebra is well adapted to multilinear calculations; to tailor it for determinant type work, we must form a graded quotient algebra. To do this, we begin with a graded submodule $S \subset M$, which is a sequence $S_i \subset M_i$, $i = 0, 1, 2, \ldots$, of submodules. Next, we make S into an ideal in M by closing it under left and right tensor product. The S of interest here is $S(M_1^2)$, which is generated by the set $\{m_1 \otimes m_1 : m_1 \in M_1\}$. Then, we form the graded module M_i/S_i, $i = 0, 1, 2, \ldots$ of quotient modules M_i/S_i, where M_o/S_o is $F[v]$ and M_1/S_1 is M_1 because S_o and S_1 are the zero module. The sequence $P_i : M_i \rightarrow M_i/S_i$ of standard quotient mappings forms a morphism $P : M \rightarrow M/S$ of graded modules. P can be made into a morphism of graded algebras by defining a product $\wedge$ in M/S such that

$$(Pm_1) \wedge (Pm_2) = P(m_1 \otimes m_2)$$

for m_1 and m_2 in M. So equipped, M/S is called the exterior algebra of M_1 and is denoted $\bigwedge(M_1)$.

$\bigwedge(M_1)$ is relevant to the pole assignment question because (5) can as well be replaced by

$$\tilde{d}_1 \wedge \tilde{d}_2 \wedge \ldots \wedge \tilde{d}_n. \tag{6}$$

APPLICATION TO POLE ASSIGNMENT

Now that we have defined a relevant exterior algebra, we can use this structure to address those specifications that are related to the position of poles in the solutions to MDP, which we stated earlier; namely, (iv) determine any fixed poles in the solutions, and (v) develop algorithms for pole assignment.

Assume that, on applying the MDP Algorithm [8], we obtained a reduced basis for the kernel:

$$\begin{bmatrix} n_1 \\ d_1 \end{bmatrix}, \begin{bmatrix} n_2 \\ d_2 \end{bmatrix}, \dots, \begin{bmatrix} n_m \\ d_m \end{bmatrix} \tag{7}$$

where $d_i \in (F[v])^n$. We desire a particular characteristic polynomial such as (6), for $n < m$. (The case $n = m$ is trivial). Without loss of generality, assume that $d_1, d_2, \dots d_m$ are the basis for a free $F[v]$-module of rank m. Then, it is possible to write [1]

$$\begin{aligned} \tilde{d}_1 \wedge \tilde{d}_2 \wedge \dots \wedge \tilde{d}_n = {} & \begin{vmatrix} f_{11} & f_{21} \dots & f_{n1} \\ \vdots & \vdots & \vdots \\ f_{1n} & f_{2n} \dots & f_{nn} \end{vmatrix} d_1 \wedge d_2 \wedge \dots \wedge d_n \\ & + \begin{vmatrix} f_{11} & f_{21} & \cdots f_{n1} \\ \vdots & \vdots & \vdots \\ f_{1(n-1)} & f_{2(n-1)} & f_{n(n-1)} \\ f_{1(n+1)} & f_{2(n+1)} \dots & f_{n(n+1)} \end{vmatrix} d_1 \wedge d_2 \wedge \dots \wedge d_{n-1} \wedge d_n \\ & + \dots + \begin{vmatrix} f_{1(m-n+1)} & f_{2(m-n+1)} \dots & f_{n(m-n+1)} \\ \vdots & \vdots & \vdots \\ f_{1m} & f_{2m} & f_{nm} \end{vmatrix} d_{m-n+1} \wedge \dots \wedge d_m \end{aligned} \tag{8}$$

where

$$\begin{aligned} \tilde{d}_1 &= f_{11} d_1 + f_{12} d_2 + \dots + f_{1m} d_m \\ &\vdots \\ \tilde{d}_n &= f_{n1} d_1 + f_{n2} d_2 + \dots + f_{nm} d_m, \end{aligned}$$

with the $n \times m$ coefficients $f_{11}, f_{12}, \dots, f_{nm}$ being ring elements.

Equation (8) defines the limits on the possible characteristic polynomials. For instance, the fact that a polynomial combination of $d_1 \wedge d_2 \wedge \dots \wedge d_n$, $d_1 \wedge d_2 \wedge \dots \wedge d_{n-1} \wedge d_{n+1}, \dots, d_{m-n+1} \wedge \dots \wedge d_m$ is an ideal generated by the GCD of the polynomials resolves issue (iv); in other words, the fixed poles in the solution to MDP are simply the roots of this GCD.

In order to address issue (v), namely, to assign additional arbitrary poles, we must be able to obtain the $n \times m$ coefficients $f_{11}, \dots, f_{nm}$. If we rewrite (8), with a change of notation, as

$$\begin{aligned} \tilde{d}_1 \wedge \tilde{d}_2 \wedge \dots \wedge \tilde{d}_n = {} & g_k(d_1 \wedge d_2 \wedge \dots \wedge d_n) + g_{k-1}(d_1 \wedge \dots \wedge d_{n-1} \wedge d_{n+1}) \\ & + \dots + g_1(d_{m-n+1} \wedge \dots \wedge d_m). \end{aligned} \tag{9}$$

where $k = m!/[n!(m-n)!]$, and $g_1,\ldots, g_k$ are ring elements, we can solve for coefficients $g_1,\ldots g_k$ in (9) if a solution to (9) exists. So the main problem concerns the method by which the $f_{11},\ldots, f_{nm}$ may be obtained, once $g_1,\ldots, g_k$ are known.

The power of the exterior algebra is aptly demonstrated at this point, for it renders the approach to the solution transparent. Simply project the elements $\tilde{d}_1,\ldots, \tilde{d}_n$ into the rank (n+1) module spanned by $d_1, d_2,\ldots, d_{n+1}$. Then, the corresponding projections (which we call $\underline{\tilde{d}}_1,\ldots, \underline{\tilde{d}}_n$) can be expressed as

$$\begin{aligned} \underline{\tilde{d}}_1 &= f_{11}\, d_1 + f_{12}\, d_2 + \ldots + f_{1(n+1)}\, d_{n+1} \\ &\vdots \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ \underline{\tilde{d}}_n &= f_{n1}\, d_1 + f_{n2}\, d_2 + \ldots + f_{n(n+1)}\, d_{n+1}. \end{aligned}$$

Corresponding to (9), $\underline{\tilde{d}}_1 \wedge \underline{\tilde{d}}_2 \wedge \cdots \wedge \underline{\tilde{d}}_n$ will be expressed exclusively in terms of those n + 1 determinant terms in the right member of (9), that involve $d_1,\ldots, d_{n+1}$ exclusively. The appropriate (n+1) coefficients g_i are already known, say $g'_{n+1}, g'_n,\ldots, g'_1$, and we now have the equation

$$\begin{aligned} \underline{\tilde{d}}_1 \wedge \underline{\tilde{d}}_2 \wedge \ldots \wedge \underline{\tilde{d}}_n &= g'_{n+1}(d_1 \wedge d_2 \wedge \cdots \wedge d_n) + \hat{g}_n(d_1 \wedge \cdots \wedge d_{n-1} \wedge d_{n+1}) \\ &+ g'_{n-1}(d_1 \wedge d_2 \wedge \cdots \wedge d_{n-2} \wedge d_n \wedge d_{n+1}) + \ldots + g'_1(d_2 \wedge d_3 \wedge \cdots \wedge d_n \wedge d_{n+1}) \end{aligned} \tag{10}$$

From the fact that $(\underline{\tilde{d}}_1 \wedge \underline{\tilde{d}}_2 \wedge \cdots \wedge \underline{\tilde{d}}_k \wedge \cdots \wedge \underline{\tilde{d}}_n) \wedge \underline{\tilde{d}}_k$, $k = 1, 2,\ldots, n$, equals zero, it is easy to show that

$$\begin{bmatrix} f_{11} \\ f_{12} \\ \vdots \\ f_{1(n+1)} \end{bmatrix}, \quad \begin{bmatrix} f_{21} \\ f_{22} \\ \vdots \\ f_{2(n+1)} \end{bmatrix}, \quad \ldots, \quad \begin{bmatrix} f_{n1} \\ f_{n2} \\ \vdots \\ f_{n(n+1)} \end{bmatrix}$$

must be in $\mathrm{Ker}[g'_1 \ \ -g'_2 \ \ g'_3 \ldots (-1)^n g'_{n+1}]$. Since this kernel has rank n, candidates for f_{ij}, $i = 1, 2,\ldots, n$, $j = 1, 2,\ldots, n + 1$ are easily found.

If m equals n + 1, the question of whether (6) is obtainable has an algorithmic answer following the ideas above, and an example is provided in the following section. If m exceeds n + 1, the ideas can be applied $m!/[(n+1)!(m-n-1)!]$ times to yield that many kernel constraints on the f_{ij}. Because it is known [11] that elements in M_n/S_n need not always be decomposable (i.e., in terms of the $\tilde{d}_i$), it is not to be expected that solutions for the f_{ij} will always exist. Nonetheless, the approach described here does lay out the constraints which then allow a calculation of the possible f_{ij}'s in a straightforward manner.

EXAMPLE [8]

Let F = R, the real field, and let

$$G_1(v) = \frac{1}{v^2 + 3v + 2} \begin{bmatrix} v + 1 & v + 3 & v(v+3) \\ v + 2 & v(v+2) & 0 \end{bmatrix}$$

In this example, we discuss a version of the problem (2) presented originally by Wang and Davison [3] and by Forney [7], where $G_2(v)$ is the identity. In this case, therefore, equation (2) becomes

$$[N_1 \vdots -d_1(v)I] \begin{bmatrix} N \\ D \end{bmatrix} = [0]. \tag{11}$$

We express G_1 as

$$\begin{bmatrix} v^2 + 3v + 2 & 0 \\ 0 & v + 1 \end{bmatrix}^{-1} \begin{bmatrix} v + 1 & v + 3 & v(v+3) \\ 1 & v & 0 \end{bmatrix}$$

and rewrite (11) as

$$\begin{bmatrix} v + 1 & v + 3 & v(v+3) & -v^2 - 3v - 2 & 0 \\ 1 & v & 0 & 0 & -v - 1 \end{bmatrix} \begin{bmatrix} N \\ D \end{bmatrix} = [0]$$

Our map T of free $F[v]$-modules $(R[v])^5$ to $(R[v])^2$ is thus defined by the matrix

$$T = \begin{bmatrix} v + 1 & v + 3 & v(v+3) & -v^2 - 3v - 2 & 0 \\ 1 & v & 0 & 0 & -v - 1 \end{bmatrix}.$$

From the last two columns of T, we see that

$$r(\text{Im } T) = 2,$$

and, thus,

$$r(\text{Ker } T) = 5 - r(\text{Im } T) = 3.$$

Applying Step 1 of the MDP Algorithm [8] yields a basis b_1', b_2', b_3' for Ker T given by

$$\begin{bmatrix} 0 \\ v(v+1) \\ -(v+1) \\ 0 \\ v^2 \end{bmatrix}, \quad \begin{bmatrix} v + 2 \\ -v(v+2) \\ v + 2 \\ 1 \\ (-v+1)(v+2) \end{bmatrix}, \quad \begin{bmatrix} .5v + 1.5 \\ .5v^3 - 2v - .5 \\ -.5v^2 + 1.5 \\ 0 \\ (v-1)(.5v^2-1.5) \end{bmatrix}.$$

Next, applying Step 2 of the MDP Algorithm, we obtain a reduced basis b_1, b_2, b_3 of the form

$$\begin{bmatrix} v + 2 \\ -v \\ 1 \\ 1 \\ -v + 2 \end{bmatrix}, \quad \begin{bmatrix} -1 \\ 2v + 1 \\ -1 \\ 1 \\ 2v - 1 \end{bmatrix}, \quad \begin{bmatrix} v + 6 \\ -7v - 2 \\ v + 6 \\ v \\ -7v + 6 \end{bmatrix}. \tag{12}$$

We could continue with the MDP Algorithm and complete the minimal inverse system design. For purposes of brevity, however, we stop the algorithm here, and focus attention, instead, on obtaining a desired characteristic polynomial, say

$$|D| = v^2 + 3v + 4.$$

Let us call the elements formed from the last two rows in (12) d_1, d_2 and d_3 respectively:

$$d_1 \triangleq \begin{bmatrix} 1 \\ -v + 2 \end{bmatrix} \qquad d_2 \triangleq \begin{bmatrix} 1 \\ 2v - 1 \end{bmatrix} \qquad d_3 \triangleq \begin{bmatrix} v \\ -7v + 6 \end{bmatrix}.$$

The first step is to find g_1, g_2 and g_3 to satisfy

$$\begin{aligned} v^2 + 3v + 4 &= g_3(d_1 \wedge d_2) + g_2(d_1 \wedge d_3) + g_1(d_1 \wedge d_3) \\ &= g_3(3v-3) + g_2(v^2-9v+6) + g_1(-2v^2-6v+6). \end{aligned}$$

Solving this equation, we obtain

$$(g_3 \quad g_2 \quad g_1)^T = (-28/3 \quad -7/3 \quad -5/3)^T.$$

Next, we solve for a basis $(f_{11} \quad f_{12} \quad f_{13})^T$ and $(f_{21} \quad f_{22} \quad f_{23})^T \in \mathrm{Ker}$ $(g_1 \quad -g_2 \quad g_3)$. This gives

$$[f_{11} \quad f_{12} \quad f_{13}] = [7/5 \quad 1 \quad 0],$$

and

$$[f_{21} \quad f_{22} \quad f_{23}] = [-28/5 \quad 0 \quad 1].$$

Finally, $\tilde{D}$ is obtained as

$$\tilde{D} = [\tilde{d}_1 \quad \tilde{d}_2]$$

where

$$\tilde{d}_1 = 7/5 \begin{bmatrix} 1 \\ -v + 2 \end{bmatrix} + 1 \begin{bmatrix} 1 \\ 2v - 1 \end{bmatrix} = \begin{bmatrix} 12/5 \\ 3v/5 + 9/5 \end{bmatrix} \tag{13A}$$

and

$$\tilde{d}_2 = -28/5 \begin{bmatrix} 1 \\ -v + 2 \end{bmatrix} + 1 \begin{bmatrix} v \\ -7v + 6 \end{bmatrix} = \begin{bmatrix} v - 28/5 \\ -7v/5 - 26/5 \end{bmatrix} \tag{13B}$$

To check, we calculate

$$|\tilde{D}| = \begin{vmatrix} 12/5 & v - 28/5 \\ 3v/5 + 9/5 & -7v/5 - 26/5 \end{vmatrix} = -3/5(v^2+3v+4)$$

The additional scalar factor was introduced because the first column was scaled by -3/5 while reducing to echelon form; this can be trivially corrected by scaling the second column of $\tilde{N}$ and $\tilde{D}$ by -5/3, to obtain the desired solution as

$$\begin{bmatrix} 7v/5 + 9/5 & 23v/3 + 26/3 \\ 3v/5 + 1 & 7v/3 + 10/3 \\ 2/5 & -5v/3 - 2/3 \end{bmatrix} \begin{bmatrix} 12/5 & -5v/3 + 28/3 \\ 3v/5 + 9/5 & 7v/3 + 26/3 \end{bmatrix}^{-1}$$

The matrix $\tilde{N}$ is found, of course, by applying the same linear combination of (13) to the upper portions of the elements in (12).

REMARKS

Some care is advisable in selecting polynomial candidates for $|\tilde{D}|$. We have already mentioned the necessity of including fixed poles in any such candidate. Of equal importance is the possibility of making a linear dynamical interpretation for $\tilde{N}\tilde{D}^{-1}$. Here the solution to MDP, which determines candidates D such that $|D|$ has least degree, supplies a fundamental guideline. It is not possible to specify a $|\tilde{D}|$ of degree less than this minimal degree and still obtain a realizable $\tilde{N}\tilde{D}^{-1}$. A reasonable approach to this question is to specify $|\tilde{D}|$ in such a way that its degree is equal to the minimal degree determined from the MDP algorithm.

A number of very interesting research questions remain open in the approach which has been described above. As presented, the conditions on f_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, are necessary only, and not sufficient. Even when $m = n + 1$, when the candidates determined are the only ones, a check such as that performed in the example appears to be prudent. (If it were possible to restrict the problem to an F-vector space structure, the conditions on f_{ij}, when $m = n + 1$, do become sufficient as well. Such a restriction, however, seems to be at the expense of enrichment in the set of possible $|\tilde{D}|$'s). When $m > n + 1$, additional work appears relevant in the area of determining whether or not the necessary kernel restrictions on the f_{ij} permit any solutions compatible with the necessary conditions.

The structure afforded by appropriate exterior algebras in this type of problem offers promise for settlement of some of these issues, and ensuing investigations are now in progress.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation under Grants GK 37285 and ENG 75 22322.

REFERENCES

1. M. K. Sain, "The Growing Algebraic Presence in Systems Engineering: An Introduction," Proceedings of the IEEE, Vol. 64, No. 1, pp. 96-111, January 1976.

2. M. K. Sain, "Applications of Modern Algebra in Engineering--Introductory Lecture Notes," Technical Report No. EE-752, Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana, 1975.

3. S. H. Wang and E. J. Davison, "A Minimization Algorithm for the Design of Linear Multivariable Systems," IEEE Transactions on Automatic Control, Vol. 18, No. 3, pp. 220-225, June 1973.

4. H. H. Rosenbrock, State Space and Multivariable Theory, New York: Wiley, 1970.

5. V. M. Popov, "Some Properties of Control Systems with Irreducible Matrix Transfer Functions," Seminar on Differential Equations and Dynamical Systems, (Lecture Notes in Mathematics, No. 144), pp. 250-261, New York: Springer, 1970.

6. W. A. Wolovich, "The Determination of State-Space Representations for Linear Multivariable Systems," Automatica, Vol. 9, pp. 97-106, January 1973.

7. G. D. Forney, "Minimal Bases of Rational Vector Spaces, with Applications to Linear Multivariable Systems," SIAM Journal on Control, Vol. 13 pp. 493-520, May 1975.

8. M. K. Sain, "A Free-Modular Algorithm for Minimal Design of Linear Multivariable Systems," Proc. Sixth IFAC World Congress, Part 1, pp. 9.1-1--9.1-7, Boston, August 1975.

9. R. R. Gejji, "A Computer Program to Find the Kernel of a Polynomial Operator," to be presented at the Fourteenth Annual Allerton Conference on Circuit and System Theory, 1976.

10. V. Seshadri, "Compensation of Multivariable Control Systems," M.S. Thesis, Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana, May 1976; also available as Technical Report No. EE-763, May 9, 1976.

11. S. MacLane and G. Birkhoff, Algebra, London: Macmillan, 1967.

SAMPLED-DATA OBSERVABILITY OF CONTINUOUS-TIME SYSTEMS

ROBERT A. SCHLUETER
Department of Electrical Engineering
and Systems Science
Michigan State University
East Lansing, Michigan 48824

RICHARD K. T. MA
IBM Thomas J. Watson
Research Center
Yorktown Heights
New York 10598

ABSTRACT

Sampled-data observability is defined for continuous-time systems where the number of samples and the sequence of sampling times at which measurements are made are almost arbitrary. A continuous-time system is proved to be observable with some finite sequence of sampled output measurements if and only if it is observable with continuous output measurements. Necessary and sufficient conditions for observability are also derived for the special case where the number of samples is specified, but the sampling times can be selected almost arbitrarily. This theorem indicates restrictions on the number of samples used, the selection of the sampling times sequence, and the structure of the continuous-time system. A procedure for testing observability of a system on a particular sequence of sampling times is then developed based on these theoretical results.

1. INTRODUCTION

Controllability and observability were originally developed as purely mathematical concepts. However, they were soon found to be related to the possibility of achieving a desired degree of control and obtaining the desired information about the system.

Controllability assures that the optimal control law designed for a linear system using a quadratic performance index will be asymptotically stable. Observability assures the Kalman filter will be asymptotically stable. Moreover, controllability and observability are also important in the realm of mathematical modeling. Although a state space model is desired for analytic design of the control law, one often starts with an input-output model obtained experimentally. The minimal realization, which does not introduce any phenomena that cannot be accounted for by an input-output description of the system, is intimately related to the concepts of controllability and observability. Thus, controllability and observability are important concepts in the areas of control, estimation, and identification of dynamic systems.

Controllability and observability of sampled data systems are well defined concepts if the number and lengths of sampling times are specified apriori. Necessary and sufficient conditions for controllability or observability are well known [1, 2] and have been used for testing controllability or observability on a particular sampling times sequence. The theory is not well developed for the case where the number and lengths of sampling times can be selected almost arbitrarily. The importance of extending the present theory to the general case has become apparent with the formulation of optimal measurement scheduling problems in control [3-6], state estimation [7], and identification [8]. Necessary and sufficient conditions for observability and controllability, which place requirements on the structure of the continuous-time system, the number of samples, and the lengths of sampling intervals, would indicate whether the system could be observable and/or controllable on an almost arbitrary sequence of sampling times. These necessary and sufficient conditions could also be used to improve the efficiency and convergence properties of computational algorithms developed for solving these optimal scheduling problems. These rules can also assist system designers in selecting sampling rates or aperiodic sampling criteria that will ultimately improve the overall system performance.

Some results on controllability and observability have been derived by Troch [9] and Gilchrist [10] for the case where the number of samples are specified, but the sampling times can be selected almost arbitrarily. A continuous-time system was defined as being p-sampled data observable if there existed a sequence of p sampling times for which the state could be determined from the sampled output measurements and the control history [10]. A sufficient condition for q-sampled data observability was derived for an observable continuous-time system with minimal polynomial of order q [9]. This condition required that a matrix function of the q sampling times be nonsingular. A lemma was then proved which stated that this matrix function would be nonsingular for all sequences of q sampling if the eigenvalues were all real and for all sequences belonging to a particular open bounded interval if some eigenvalues were complex. These results guaranteed the existence of q sampling times which would make a continuous-time system, observable with continuous output measurements, observable with q sets of sampled output measurements. Although these results are significant, they are only sufficient conditions and are restricted to the case where $p = q$. Moreover, no relationships between sampling times and system eigenvalues, which can cause loss of observability for an otherwise observable system, were derived.

Three sets of necessary and sufficient conditions for p-sampled data observability are derived in section 2 of the paper; these are not restricted to the case where $p = q$. A corollary is then proven that extends Troch's [9] sufficient condition to the case where $p \geq q$. Observability is defined in section 3 for continuous-time systems where both the number of samples and the lengths of sampling intervals can be selected almost arbitrarily. A theorem is then proven which states that a finite sampling times sequence can always be found that will permit calculation of the state from the finite set of sampled output measurements if and only if the system is observable with continuous output measurements.

2. NECESSARY AND SUFFICIENT CONDITIONS FOR P-SAMPLED-DATA OBSERVABILITY

The observability of continuous-time systems, which are described by equations

$$\begin{aligned} \dot{\underline{x}}(t) &= \underline{A}\,\underline{x}(t) + \underline{B}\,\underline{u}(t) \\ \underline{y}(t) &= \underline{C}\,\underline{x}(t) \end{aligned} \tag{1}$$

where $\underline{x} \in R^n$, $\underline{u} \in R^r$, $\underline{y} \in R^m$, and $\underline{A}$, $\underline{B}$, $\underline{C}$ are constant compatible matrices, is investigated for the case where the output $\underline{y}(t)$ is sampled at an almost arbitrary sequence of sampling times, satisfying

$$0 < t_0 < t_1 < \ldots < t_{N-1}. \tag{2}$$

Observability is investigated in this section for the case where the number of samples is specified $(N = p)$ and in the next section for the case where N is arbitrary. The following definition of p-sampled-data observability is taken from Troch [9]. The concept "completely observable" shall be used as defined by Kalman [11].

Definition 1. System (1) is completely p-sampled-data observable if there exists p-sampling times $\{t_i\}_{i=0}^{p=1}$ that satisfy (2) such that any initial state $\underline{x}(0)$ can be calculated from the output at the sampling times $\{\underline{y}(t_i)\}_{i=0}^{p-1}$ and control history $\underline{u}_{[0,t_{p-1}]}$.

The following theorem is obtained by extending a theorem for systems where the number and length of sampling intervals are specified.

Theorem 1. System (1) is completely p-sampled-data observable if and only if there exists p-sampling times $\{t_i\}_{i=0}^{p-1}$ that satisfy (2) for which

$$\underline{Q}_o(t_0, t_1, \ldots, t_{p-1}) = [\underline{\Phi}_0 \underline{C}^T \,\vdots\, \underline{\Phi}_1^T \underline{C}^T \,\vdots\, \underline{\Phi}_2^T \underline{C}^T \,\vdots\, \ldots \,\vdots\, \underline{\Phi}_{p-1}^T \underline{C}^T] \tag{3}$$

has rank equal to n where

$$\underline{\Phi}_i = \underline{\epsilon}^{At_i}.$$

Troch [9] has derived a sufficient condition for p-sampled-data observability for observable continuous-time systems (1) in terms of a condition on the sampling times. This sufficient condition holds for $p = q$, where q is the order of the minimal polynomial for system matrix $\underline{A}$ and where the fundamental matrix can be expressed as

$$\underline{\epsilon}^{At} = \sum_{j=0}^{q-1} \alpha_j(t)\,\underline{A}^j .$$

Theorem 2. Suppose that system (1) is completely observable in the sense of Kalman [11]. Then (1) is completely q-sampled-data observable if a sequence of q sampling times $\{t_i\}_{i=0}^{q-1}$ can be found that make the $q \times q$ matrix

$$\{\underline{X}_o(t_0, t_1, \ldots, t_{q-1})\}_{ji} = \alpha_{j-1}(t_{i-1})\quad i, j = 1, 2, \ldots, q \tag{4}$$

nonsingular.

Troch [9] has also established the following lemma that establishes that a sequence $\{t_i\}_{i=0}^{q-1}$ can always be found for which $\underline{X}_o$ is nonsingular.

Lemma 1. Given system (1) with maximum eigenfrequency ω (i. e., the greatest imaginary part of the eigenvalues of $\underline{A}$), the matrix $\underline{X}_o(t_0, t_1, \ldots, t_{q-1})$ is nonsingular for all sequence $\{t_i\}_{i=0}^{q-1}$ that satisfy

$$0 \le t_1 < t_2 < \ldots < t_{q-1} < \frac{\pi}{\omega} . \tag{5}$$

If all eigenvalues are real, then $\underline{X}_o(t_0, t_1, \ldots, t_{q-1})$ is nonsingular for all sequences that satisfy

$$0 \le t_0 < t_1 < t_2 < \ldots < t_{q-1} < \infty . \tag{6}$$

The inequality (5) can be used in the formulation of measurement scheduling problems [3-8] so that observability will be preserved on all feasible sampling times sequences. This constraint need only be imposed if the system has complex eigenvalues.

The results of Theorem 2 and Lemma 1 indicate that the system will always be q-sampled-data observable if it is observable with continuous output measurements. These results do not indicate necessary conditions for p-sampled-data observability and are restricted to the very special case where $p = q$. Necessary and sufficient conditions, which place constraints on the structure of system (1), the number of samples, and the selection of a sampling times sequence, can be used in determining whether an optimal measurement scheduling problem [3-8] is well posed for a particular system using a specified number of samples. The following theorem will now state such a set of necessary and sufficient conditions for p-sampled-data observability.

Theorem 3. System (1) is completely p-sampled-data observable if and only if

i) system (1) is completely observable;

ii) the number of samples satisfy $pm \geq n$;

iii) n columns of the $qm \times pm$ matrix

$$\hat{\underline{X}}_o(t_0, t_1, \ldots, t_{p-1} = \begin{bmatrix} a_0(t_0)\underline{I}, & a_0(t_i)\underline{I} & a_0(t_{p-1})\underline{I} \\ a_1(t_0)\underline{I}, & a_1(t_i)\underline{I} & a_1(t_{p-1})\underline{I} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{q-1}(t_0)\underline{I}, & a_1(t_i)\underline{I} & a_1(t_{p-1})\underline{I} \end{bmatrix} \tag{7}$$

are linearly independent and linear combinations of the rows of matrix

$$\underline{Q} = [\underline{C}^T \,|\, \underline{A}^T \underline{C}^T \,|\, \ldots \,|\, \underline{A}^{q-1^T} \underline{C}^T \,| \tag{8a}$$

for some sequence $\{t_i\}_{i=0}^{p-1}$ satisfying (2).

Proof: Necessity is proven first by noting that matrix $\underline{Q}_o(\cdot)$ can be expressed as

$$\underline{Q}_o(t_0, t_1, \ldots, t_{p-1}) = \underline{Q}\, \hat{\underline{X}}_o(t_0, t_1, \ldots, t_{p-1}) \tag{8b}$$

and from Sylvester's theorem on matrix products

$$\text{Rank}\{\underline{Q}_o(t_0, t_1, \ldots, t_{p-1})\} \leq \min(\text{Rank}\{\underline{Q}\}, \text{Rank}\{\hat{\underline{X}}_o(t_0, t_1, \ldots, t_{p-1})\}). \tag{9}$$

Matrix $\underline{Q}_o(\cdot)$ can have rank n only if matrices $\underline{Q}$ and $\hat{\underline{X}}_o(\cdot)$ have rank n, and thus conditions (i) and (ii) are necessary. Finally, the matrix product (8) can have rank n only if $\hat{\underline{X}}_o(\cdot)$ has n linearly independent columns that lie in the n dimensional subspace spanned by the qm dimensional row vectors of $\underline{Q}$ and thus (iii) is also necessary.

Sufficiency is proven by noting that if (i-iii) are satisfied and if $\{\hat{\underline{X}}_{o,i_j}\}_{j=1}^{n}$ are the n columns of $\hat{\underline{X}}_o(\cdot)$ for some $\{t_i\}_{i=0}^{p-1}$ that satisfy (iii), then the square matrix

$$\underline{Q}[\underline{X}_{o,i_1} \,|\, \underline{X}_{o,i_2} \,|\, \ldots \,|\, \underline{X}_{o,i_n}] \tag{10}$$

has rank n. Thus, (i-iii) is sufficient for $\underline{Q}_o(\cdot)$ to have rank n and the theorem is proven.

Condition (iii) states a relationship between the functions $\{a_j(t)\}_{j=0}^{q-1}$ and the rows of $\underline{Q}$ that may never be satisfied for any sequence $\{t_i\}_{i=0}^{p-1}$ if $q > p \geq n/m$. However, if $p \geq q \geq n/m$, condition (iii) can always be satisfied for some $\{t_i\}_{i=0}^{p-1}$ (Theorem 4) which will be proven from the results of the following corollary.

Corollary 1. Suppose system (1) is completely observable in the sense of Kalman. Then (1) is completely p-sampled-data observable for $p \geq q$ if a sequence of sampling times $\{t_i\}_{i=0}^{p-1}$ exists satisfying (2) for which

$$\operatorname{Rank}\{\underline{X}_o(t_0, t_1, \ldots, t_{p-1})\} = q \tag{11}$$

where

$$\{X_o(t_0, t_1, \ldots, t_{p-1})\}_{ji} = a_{j-1}(t_{i-1}) \quad \begin{array}{l} j = 1, 2, \ldots, q \\ i = 1, 2, \ldots, p. \end{array} \tag{12}$$

Proof: For $p \geq q$, matrix $\underline{Q}_o(\cdot)$ can have rank n if

$$\operatorname{Rank}\{\hat{\underline{X}}_o(t_0, t_1, \ldots, t_{p-1})\} = qm \tag{13}$$

since there would be qm linearly independent columns of $\hat{\underline{X}}_o(\cdot)$ that must span the n-dimensional subspace containing the row vectors of $\underline{Q}$. $\hat{\underline{X}}_o(\cdot)$ has rank qm if and only if condition (11) can be satisfied for some $\{t_i\}_{i=0}^{p-1}$ and therefore the corollary is proven.

Theorem 2, derived by Troch [9], is merely a special case of Corollary 1 for the case where $p = q$. However, condition (4) in Theorem 2 is quite important from a practical viewpoint. Condition (4) will not only be used in the test procedure for observability on a sequence $\{t_i\}_{i=0}^{p-1}$ developed in section 4, but will also be used to derive relationships between sampling times and system eigenvalues that can cause loss of observability in an otherwise observable system.

The following theorem states a strong necessary and sufficient condition for p-sampled-data observability for the case where $p \geq q$ and is based on the knowledge that (11) can be satisfied almost everywhere in R^p.

Theorem 4. A necessary and sufficient condition for system (1) to be completely p-sampled-data observable for $p \geq q$ is that system (1) be completely observable.

Proof: The system (1) is completely p-sampled-data observable if and only if $\underline{Q}_o(\cdot)$ has rank n for some $\{t_i\}_{i=0}^{p-1}$. Since the functions $a_j(t)$ $j = 0, 1, \ldots, q-1$ are linearly independent almost everywhere, $\hat{\underline{X}}_o(\cdot)$ and $\underline{X}_o(\cdot)$ have rank qm and q almost everywhere. Therefore, $\underline{Q}_o(\cdot)$, given in (8), has rank n almost everywhere for $p \geq q$ if and only if $\underset{\sim}{Q}$ has rank n and system (1) is p-sampled-data observable for $p \geq q$ if and only if system (1) is observable.

3. SAMPLED-DATA OBSERVABILITY

Theorems 3 and 4 indicate that observability of systems (1) with p arbitrary sampling times is impossible if $pm < n$ and cannot be guaranteed unless $p \geq q \geq n/m$. Thus, a definition of observability which specifies the number of samples can cause a sampled-data system to be called unobservable only because the number of samples selected was too small. Therefore, an extended definition of observability for sampled-data systems is proposed where both the number of samples and the sampling times can be selected almost arbitrarily. Since selection of the number of samples is as much a design decision as the selection of the sampling times themselves, the following definition is appropriate from both a theoretical and design point of view.

Definition. System (1) is completely sampled-data observable if there exists some finite number of samples (N) and some sampling times sequence satisfying (2) such that any initial state $\underline{x}(0)$ can be calculated from the outputs at the sampling times $\{\underline{y}(t_i)\}_{i=0}^{N-1}$ and control history $\underline{u}_{[o, t_{N-1}]}$.

Using this definition, the following theorem is a direct result of Theorem 4.

Theorem 5. A necessary and sufficient condition for the continuous time system (1) to be completely sampled-data observable is that system (1) be completely observable.

This theorem establishes a theoretical foundation for optimal measurement scheduling problems in control [3-6], state estimation [7], and identification [8]. The optimal linear regulator with optimal periodic, aperiodic, and adaptive measurement schedules [3-6] is not well posed unless the continuous-time system is sampled-data observable. The estimation [7] and identification [8] problems can be posed even though the system is not completely sampled-data observable, but the problem objective could not be fully accomplished in this case. This theorem indicates that observability with continuous output measurements is necessary and sufficient to guarantee observability with some finite set of sampled output measurements and thus the theorem provides essential information before a solution of an optimal measurement scheduling problem is attempted. The theory presented also indicates a system with complex eigenvalues will not be observable with all measurement schedules even if it is sampled-data observable. However, the optimal measurement schedule for the state estimation [7] and identification [8] would have to make the system completely sampled-data observable if it is observable because estimation error could be minimized only if the measurement schedule made observation of every mode in the system possible. The performance index for these estimation problems [7, 8] would have local maxima at points $(t_0, t_1, \ldots, t_{p-1}) \in R^p$ at which the system was unobservable since the system would be observable for almost all $(t_0, t_1, \ldots, t_{p-1}) \in R^p$ for $p \geq q$.

In the previous definitions of observability for a sampled-data system [1, 10], either the number of sampling intervals and the length of each sampling interval are both specified or the number of sampling intervals is specified and the lengths of sampling intervals are considered design parameters. In both of these definitions, a system which is observable with continuous measurements may not be observable with sampled measurements. This extended definition, where both the number of sampling intervals and the lengths of sampling intervals are design parameters, permits the preservation of observability when the outputs are no longer measured continuously but are sampled. This preservation of observability under the imposition of sampling requires a system designer to view the sensor and its sampling intervals sequence, specified by

$$\{t_i\}_{i=0}^{N-1} \quad \text{and} \quad N$$

to be part of the measurement-control system rather than part of the plant being observed. This perspective is required since the sampling intervals sequence are design parameters in this extended definition of observability.

The formulation of the measurement scheduling problem in control [3-6], state estimation [7], and identification [8] indicates recognition that at least the sampling times sequence $\{t_i\}_{i=0}^{N-1}$, if not the number of samples (N), should be considered design parameters. However, in the general formulation of these problems, the performance index was identical to the performance index used for the same problem when output measurements were made continuously, indicating that the measurement schedule should be designed strictly on performance without consideration for cost.

4. CONCLUSIONS

Necessary and sufficient conditions for p-sampled-data observability were derived which required that (1) the number of sampled output measurements (pm) be greater or equal to the order of the system; (2) the system be observable with continuous output measurements; and (3) the p-sampling exist so that n columns of a matrix $\hat{\underline{X}}_o(t_0, t_1, \ldots, t_{p-1})$ be linearly independent and linear combinations of the row vectors of $\underline{Q}$. A corollary was proven that established a sufficient condition for p-sampled-data observability ($p \geq q$) for an observable system is then proven which requires existence of a sequence $\{t_i\}_{i=0}^{p-1}$ that matrix $\underline{X}_o(t_0, t_1, \ldots, t_{p-1})$ have rank q. Troch's sufficient condition for q-sampled-data observability (Theorem 2) appears as a special case of this corollary for $p = q$. Finally, a strong necessary and sufficient condition for p-sampled-data observability ($p \geq q$) is shown to require that the continuous-time system be observable with continuous output measurements.

Sampled-data observability is then defined for continuous-time systems where both the number of samples and the length of sampling intervals can be selected almost arbitrarily. A necessary and sufficient condition for sampled-data observability is then shown to only require that the continuous-time system be observable with continuous output measurements.

5. REFERENCES

1. Sarachik, P.E., and E. Kreindler. "Controllability and observability of linear discrete time systems," Intl. J. Control, 419-432 (May 1965).

2. Kalman, Ho, and Narendra. Contributions to Differential Equations, Vol. I, 189-213 (1962).

3. Schlueter, R.A. "The optimal linear regulator with constrained sampling times," IEEE Trans. on Automatic Control AC-18, 515-518 (October 1973).

4. Schlueter, R.A., and A.H. Levis. "The optimal linear regulator with state dependent sampling," IEEE Trans. on Automatic Control AC-18, 512-515 (October 1973).

5. Schlueter, R.A., and A.H. Levis. "The optimal adaptive sampled-data regulator," ASME J. of Dynamic Systems, Measurement, and Control 96G: 3, 334-340 (September 1974).

6. Ma, R.K.T., and R.A. Schlueter. "Optimal control system design: the sampling and control problem," 1976 IEEE Decision and Control Conf. (December 1976).

7. Mehra, R.K. "Optimization of measurement schedules and sensor designs for linear dynamic systems," IEEE Trans. on Automatic Control AC-21, 55-64 (February 1976).

8. Goodwin, G.C., M.B. Zarrop, and R.L. Payne. "Coupled design of test signals, sampling intervals, and filters for system identification," IEEE Trans. on Automatic Control AC-19, 748-752 (December 1974).

9. Troch, I. "Sampling with arbitrary choice of sampling instants," Automatica 9, 117-124 (1973).

10. Gilchrist, J.D. "n-observability for linear systems," IEEE Trans. on Automatic Control AC-11, 388-395 (July 1966).

11. Kalman, R.E. "On the general theory of control systems," Proc. of 1st Congress of IFAC, 481-492 (1961).

LINEAR SYSTEM ESTIMATES VIA INTERMEDIATE DOMAIN PROCESSING

SORIN COHN-SFETCU
Bell-Northern Research
Ottawa, Canada

ABSTRACT

This paper addresses the problem of determining the characteristics of a linear system from an estimate of an intermediate model, which is linear and invariant to shifts governed by abstract group operations.

INTRODUCTION

It has been shown recently[1,2] that it is possible and computationally advantageous to estimate the secondary statistics of stationary data via an intermediate domain processing. It should be also possible to estimate linear systems via estimates of intermediate models. The study of this problem was initiated by Pearl[3], who analyzed the particular case of time-invariant systems and dyadic Walsh intermediate processing in a finite discrete case ($N=2^n$ points).

This paper generalizes previous theoretical results in two directions. First, the intermediate model can be calculated with respect to any unitary transformation representation which has an abelian group structure. Secondly, the question of recovering the original system from its intermediate model is generalized to systems which are not necessarily time-invariant and symmetric or causal.

FORMULATION OF THE PROBLEM

Let us consider a linear system as a triplet $\{X,Y,\Gamma\}$, where the input and output sets X and Y are the space $K[S]$ of K-valued functions on a set S, K being a field (possibly the complex number field C). The binary relation Γ is represented by a bounded linear operator $\mathcal{H}$. Translation invariance preconditions an abelian group structure for S, the functions' domain; otherwise, the function space $K[S]$ would not be closed under the operation of translation. For time-invariant systems, S must be either the group of real numbers or that of integers, under the operation of arithmetic addition.

If S is a finite discrete set of N points, the input-output (I/O) relationship takes a matrix form

$$\underline{y} = H\,\underline{x} \tag{1}$$

where H is the N×N system matrix and $\underline{y}$ and $\underline{x}$ are N-dimensional vectors. It is customary to say that a linear time-invariant (LTI) system is characterized by a matrix

$$(H_T)_{i,j} = h_T(i-j); \qquad i,j = 0,1,\ldots, N-1, \tag{2}$$

although true time-invariance is achieved only asymptotically, as $N\to\infty$. Figure 1a presents H_T for N=9.

$$H_T = \begin{array}{ccccccccc}
h_T(0) & h_T(-1) & h_T(-2) & h_T(-3) & h_T(-4) & h_T(-5) & h_T(-6) & h_T(-7) & h_T(-8) \\
h_T(1) & (0) & (-1) & (-2) & (-3) & (-4) & (-5) & (-6) & (-7) \\
h_T(2) & (1) & (0) & (-1) & (-2) & (-3) & (-4) & (-5) & (-6) \\
h_T(3) & (2) & (1) & (0) & (-1) & (-2) & (-3) & (-4) & (-5) \\
h_T(4) & (3) & (2) & (1) & (0) & (-1) & (-2) & (-3) & (-4) \\
h_T(5) & (4) & (3) & (2) & (1) & (0) & (-1) & (-2) & (-3) \\
h_T(6) & (5) & (4) & (3) & (2) & (1) & (0) & (-1) & (-2) \\
h_T(7) & (6) & (5) & (4) & (3) & (2) & (1) & (0) & (-1) \\
h_T(8) & (7) & (6) & (5) & (4) & (3) & (2) & (1) & (0)
\end{array}$$

(a)

$$H_A = \begin{array}{ccccccccc}
h_A(0) & h_A(2) & h_A(1) & h_A(6) & h_A(8) & h_A(7) & h_A(3) & h_A(5) & h_A(4) \\
h_A(1) & (0) & (2) & (7) & (6) & (8) & (4) & (3) & (5) \\
h_A(2) & (1) & (0) & (8) & (7) & (6) & (5) & (4) & (3) \\
h_A(3) & (5) & (4) & (0) & (2) & (1) & (6) & (8) & (7) \\
h_A(4) & (3) & (5) & (1) & (0) & (2) & (7) & (6) & (8) \\
h_A(5) & (4) & (3) & (2) & (1) & (0) & (8) & (7) & (6) \\
h_A(6) & (8) & (7) & (3) & (5) & (4) & (0) & (2) & (1) \\
h_A(7) & (6) & (8) & (4) & (3) & (5) & (1) & (0) & (2) \\
h_A(8) & (7) & (6) & (5) & (4) & (3) & (2) & (1) & (0)
\end{array}$$

Figure 1. Linear system matrices for a discrete domain of order 9:
a) time invariant system matrix (2N-1 distinct elements)
b) 3-adic invariant system matrix (N distinct elements).

Note: q-adic addition between positive integers can be defined as follows:

$$x,y \in P_{q^n} = \{0,1,\ldots,q^n-1\}$$

$$x \triangleq \sum_{j=0}^{n-1} x_j q^j; \quad x_j \in P_q = \{0,1,\ldots,q-1\}$$

$$x \dotplus y \triangleq \sum_{j=0}^{n-1} q^j (x_j+y_j)_{\bmod q}.$$

If S is a (locally compact) abelian group, G=(S,$\overset{\circ}{+}$),under some abstract group operation $\overset{\circ}{+}$ (for instance, addition modulo N, or q-adic addition[4,5] if $N=q^n$), then a system is said to be linear a-invariant (LAI) if its I/O relationship is described by an a-convolution, which takes the matrix form (1) with a system matrix characterized by

$$(H_A)_{i,j} = h_A(i\overset{\circ}{-}j); \quad h_A : G \to K; \quad i,j \in G \tag{3}$$

Figure 1b illustrates the 3-adic invariant system matrix for N=9.

Let F denote the Fourier transform[6] in the space K[G]. The corresponding transformation matrix F is unitary and has as rows (columns) the characters χ_m, $m\varepsilon G^*$ (G^* = dual group of G, and it is isomorphic to G if G is finite), such that

$$\begin{aligned} &F_{m,n} = F_{n,m} \qquad F_{o,m} = N^{-1/2} \qquad F_{m,n}F_{m,p} = N^{-1/2}F_{m,n\overset{\circ}{+}p} \\ &(F^{-1})_{m,n} = F_{m,\overset{\circ}{-}n} \qquad \sum_{k\varepsilon G} F_{k,m} = N^{1/2}\delta_{k,m} = \begin{cases} N^{1/2} & k=m \\ 0 & k\neq m \end{cases} \end{aligned} \tag{4}$$

Any matrix having the structure (3) is diagonalized by a similarity transformation $F^{-1}(\cdot)F$. Let $D_A=F^{-1}\cdot H_A\cdot F$, with $(D_A)_{i,j}=d_i\delta_{i,j}$.

The problem we are attempting to solve consists in determining the relationship between a linear system H and its optimal a-invariant model H_A, as well as the conditions for recovering the original system H from its intermediate model estimate.

THE A-INVARIANT OPTIMAL MODEL

In a first stage, the problem of estimating a system H via an a-invariant system requires the searching of the set $\{H_A\}$ of a-invariant matrices for the element H_A° which best approximates the matrix H. A convenient criterion of optimality is that requiring the minimization of the Euclidean distance between the outputs of H_A and H over an ensemble of input signals uniformly distributed over the surface of the unit sphere. A shown by Pearl[3], this criterion reduces the problem to one of finding the element of $\{H_A\}$ which minimizes the norm $|H-H_A|$ defined by

$$|H-H_A| = [N^{-1} \sum_{i,j} (H-H_A)_{i,j}^2]^{1/2} \tag{5}$$

Theorem 1: The a-invariant system H_A° which best models a linear system H (in the sense of the above-mentioned criterion) is the system defined by

$$H_A^\circ = F\cdot(\mathrm{diag}(F^{-1}\cdot H\cdot F))\cdot F^{-1}. \tag{6}$$

The proof, which is based on the diagonal matching lemma of Pearl[7], makes use only of the unitary character of the transform F, with no reference to it as a Fourier transform. The norm (5) is rotational invariant, such that

$$|H-H_A|^2 = |F^{-1}\cdot H\cdot F-F^{-1}\cdot H_A\cdot F|^2 = N^{-1} \sum_{i,j} [(F^{-1}HF)_{i,j}-(D_A)_{i,j}]^2 \tag{7}$$

which is minimized if H°_A is chosen such that D°_A is given by the diagonal elements of $F^{-1}\cdot H\cdot F$. The residual error is determined by the off-diagonal elements of $F^{-1}\cdot H\cdot F$, which are independent of the choice of H°_A.

Attention should be drawn to the fact that no restriction has been placed on the original system $\mathcal{H}$. It may be "time-invariant", or it may be invariant to a different, b-group structure on the domain S, or it may not be invariant at all. The projection from the set $\{\mathcal{H}\}$ onto the set $\{\mathcal{H}_A\}$ is not bijective; nor is it a homomorphic mapping of linear operator algebras: the cascade of two linear systems is not modelled by cascading the respective LAI system models.

IMPULSE RESPONSE RELATIONSHIPS

The second stage of the problem concerns the relationship between the "time" representations of the system and its intermediate domain model.

Theorem 2: The relationship between the impulse response of a LTI system and its optimal LAI model is:

$$h_A(i) = N^{-1} \sum_{m=0}^{N-1} h_T(m-(m \overset{\circ}{-} i)) = N^{-1} \sum_{m=0}^{N-1} h_T((m \overset{\circ}{+} i)-i). \tag{8}$$

Similar relations can be proved to exist between the impulse responses of a b-invariant linear system and its optimal a-invariant model. The proof makes full use of the fact that $\mathcal{F}$ is a Fourier transform. Given (4), it is easy to show that (6) can be rewritten as:

$$\begin{aligned}
(H^\circ_A)_{i,j} &= \sum_k F_{i,k}(F^{-1}\cdot H\cdot F)_{k,k} F^{-1}_{k,j} \\
&= \sum_k F_{i,k}\Big(\sum_m \sum_n F^{-1}_{k,m} H_{m,n} F_{n,k}\Big) F^{-1}_{k,j} \\
&= N^{-1/2} \sum_k F_{i,k} \sum_m \sum_n H_{m,n} F^{-1}_{k,m\overset{\circ}{-}n} F^{-1}_{k,j} \\
&= N^{-3/2} \sum_m \sum_n H_{m,n} \sum_k F^{-1}_{k,m\overset{\circ}{-}n\overset{\circ}{-}i\overset{\circ}{+}j} \\
&= N^{-1} \sum_m \sum_n H_{m,n}\, \delta_{o,m\overset{\circ}{-}n\overset{\circ}{-}i\overset{\circ}{-}j}
\end{aligned} \tag{9}$$

But $h_A(i)=(H^\circ_A)_{i,o}$ and $H_{m,n}=h_T(m-n)$, so that (8) follows immediately.

RECOVERY OF ORIGINAL SYSTEM

The third stage of the problem concerns the recovery of the original system from its intermediate model. As mentioned above, the projection (6) from $\{\mathcal{H}\}$ onto $\{\mathcal{H}_A\}$ is not a one-to-one correspondence, and it is not always possible to (uniquely) obtain H from H_A. But it can be shown that:

Theorem 3: the projection $H^\circ_A(H) : \{H\} \rightarrow \{H_A\}$ defined by Th. 1 is a bijective correspondence for any linear system $\mathcal{H}$ characterized by a matrix H having at most N distinct elements.

This is simply a corollary of the fact that, since an LAI system has only N distinct elements, Th. 2 provides us with only N equations for determining the elements of H. In general, H has N^2 elements. If H is time-invariant, there are only 2N-1 distinct elements and H_T can be completely determined from its LAI model if some extra N-1 h_T-structure equations (like the symmetry or causality conditions of Pearl[3]) can be established. This theorem invalidates the assertion that one cannot recover a LTI system from its circular model (obtained via a discrete Fourier transform representation; i.e. with S a group under addition modulo N); the assertion is valid only in the absence of the additional N-1 structure conditions for $\mathcal{H}_T$.

CONCLUSIONS

A set of relations has been determined between matrices and impulse responses of linear systems and their a-invariant models. Efficient algorithms exist for identifying circular or q-adic models and for performing the inversion $H_A^{\circ} \to H$ (especially for dyadic Walsh processing), but there remains open the problem of determining the goodness (bias and variance error) of such intermediate domain estimates.

REFERENCES

1. J. Pearl, "Basis restricted transformations and performance measures for spectral representations", IEEE Trans. Inform. Theory, Vol. IT-17, pp. 751-752, 1971.

2. P.V. Lopresti and H.L. Suri, "A fast algorithm for the estimation of autocorrelation functions", IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-22, pp. 449-453, 1974.

3. J. Pearl, "Optimal dyadic models of time-invariant systems", IEEE Trans. Computers, Vol. C-24, pp. 598-603, 1975.

4. H.E. Chrestenson, "A class of generalized Walsh functions", Pacific J. Math., Vol. 5, pp. 17-31, 1955.

5. J.E. Gibbs and B. Ireland, "Some generalizations of the logical derivative", DES Rept. No. 8, National Physical Laboratory, Teddington, England, 1971.

6. P.J. Nicholson, "Algebraic theory of finite Fourier transforms", J. Comput. and Syst. Science, Vol. 5, pp. 524-547, 1971.

7. J. Pearl, "Asymptotic equivalence of spectral representations", IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, pp. 547-551, 1975.

A NORMALIZED RESIDUAL CRITERION FOR THE DETERMINATION OF STRUCTURE OF LINEAR SYSTEMS

L. C. SUEN AND R. LIU
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana 46556

ABSTRACT

This paper presents a new criterion, normalized residual criterion (NRC), for the determination of system structure from a given input/output sequence as well as estimating the standard deviation of the noise and the noise-to-output signal ratio. A comparison between the NRC and Akaike information criterion (AIC) is also discussed.

I. INTRODUCTION

Basically, system identification is to estimate the structure of the system (the order and the noncausality) as well as the coefficients of the system. Once the structure of the system can be reliably estimated, the estimation of the coefficients is straightforward. Therefore, the main problem is the estimation of the structure of the dynamical system.

The estimation of the structure of a dynamical system is especially difficult while a socio-economic system is concerned. The current method has been one of trial-and-error, within the scope of socio-economic theories and statistics theories. Box and Jenkins [1] have discussed this problem. Woodside [2] and Tse [3] proposed a criterion based on the determinant of an input-output correlation matrix. Akaike's information criterion [4-6] was developed based on the maximum likelihood estimation. Parzen [7] has used the least-squares estimation to develop his criterion based on the autoregressive transfer function.

The method to be presented here utilizes the properties of the normalized residual, and therefore it is called the normalized residual criterion (NRC). The criterion is conceptually and computationally simple, and it is found to be highly reliable.

II. THE PROBLEM

Consider a linear time-invariant n-order L-step noncausal system in a noisy environment,

$$y(t) + a_1 y(t-1) + \ldots + a_n y(t-n)$$
$$= b_{-L} u(t+L) + \ldots + b_{-1} u(t+1) + b_o u(t) + \ldots + b_n u(t-n) + v(t) \qquad (1)$$

where a_i's and b_i's are constant coefficients, with a_n and b_n not both zero and $b_{-L} \neq 0$; v(t) is the noise. The coefficients $(b_{-L},\ldots,b_{-1})$ correspond to the noncausal terms. A system is causal if L is non-positive.

Let z[i,j] denote the vector [z(i) z(i+1) ... z(j)]. When a K-length input/output sequence is given, the above equation becomes

$$y[n+1,\ K-L] = \theta(n,L)W(n,L) + v[n+1,\ K-L] \tag{2}$$

where

$$\theta(n,L) = [b_n \ldots b_{-L} \ \vdots\ -a_n \ldots -a_1]$$

$$W(n,L) = \begin{bmatrix} u(1) & u(2) & \ldots & u(K-L-n) \\ u(2) & u(3) & \ldots & u(K-L-n+1) \\ \vdots & & & \\ u(n+L+1) & u(n+L+2) & \ldots & u(K) \\ y(1) & y(2) & \ldots & y(K-L-n) \\ y(2) & y(3) & \ldots & y(K-L-n+1) \\ \vdots & & & \\ y(n) & y(n+1) & \ldots & y(K-L-1) \end{bmatrix}$$

Let

$$\theta^{o}(n,L) = y\ [n+1,\ K-L]\ W^{+}(n,L) \tag{3}$$

where W^{+} is the pseudoinverse [8] of W, and

$$z^{o}[n+1,\ K-L] = \theta^{o}(n,L)W(n,L) \tag{4}$$

$$e^{o}(n,L) = ||y\ [n+1,\ K-L] - z^{o}\ [n+1,\ K-L]\ ||^{2}/||y\ [n+1,\ K-L]||^{2} \tag{5}$$

where $e^{o}(n,L)$ is called the normalized residual, which can be calculated from the input/output data for a given pair (n,L).

Definition 1. Let

$$\delta(n,L) = ||v\ [n+1,\ K-L]\ ||/||y\ [n+1,\ K-L]||$$

be the noise-to-output signal ratio.

In this paper, we study the properties of the normalized residual and then apply them to the determination of the system structure as well as the estimation of the standard deviation of the noise and δ.

III. THE MAIN RESULTS

In this paper, we assume that the noise v(t) is zero-mean with covariance $E[v(i)v(j)] = \sigma^{2}\delta_{ij}$, where E[x] is the expectation of x.

Let [θ(n,L)] denote the system whose parameters are θ(n,L).

Assertion 1. If [θ(n,L)] satisfies (1), then

$$E[||y[n+1,\ K-L]||^{2}e^{o}(n,L)|W(n,L)] = (K-3n-2L-1)\sigma^{2}$$

where E[x|y] the conditional expectation of x given y.

In view of the above assertion and definition 1, σ and δ can be estimated.

Criterion 1. The estimation of σ and δ is given by

$$\hat{\sigma}^2 = ||y[n+1,\ K-L]||^2\ e^o(n,L)/(K-3n-2L-1)$$

$$\hat{\delta}^2 = [(k-n-L)/(K-3n-2L-1)]\ e^o(n,L)$$

Criterion 1 allows us to estimate the variance of the noise and the noise-to-output signal ratio, if the order n of the system and the noncausality L are known. Next, the estimation of n and L will be considered.

It can be shown that $e^o(n,L)$ can be represented by the following approximation

$$e^o(n,L) \doteq \Delta^o(n,L) + \frac{K-3n-2L-1}{K-n-L}\ \delta^2 \tag{6}$$

where

$$\Delta^o(n,L) = \frac{||y[n+1,\ K-L] - \theta^o(n,L)W(n,L) - v[n+1,\ K-L]Q(n,L)||^2}{||y[n+1,\ K-L]||^2}$$

and

$$Q(n,L) = I - W^+(n,L)W(n,L)$$

Assertion 2. Let an input/output sequence be generated from a system $[\theta(n^o,L^o)]$. Then

$$\Delta^o(n,L) = 0 \quad \text{for } n \geq n^o \text{ and } L \geq L^o.$$

Let the autocorrelation sequence of u(t) be defined by

$$R_u(\tau) = E[u(t)u(t+\tau)]$$

and let $h^o(E)$ be a minimal polynomial for which $h(E)R_u(\tau) = 0$, where E is a shift-operator, $Ex(t) = x(t+1)$. Let d be the degree of $h^o(E)$.

Assertion 3. Let an input/output sequence be generated from a minimal degree system $[\theta(n^o,L^o)]$. Assume that $E[u(t-i)v(t-j)] = 0$ for all i,j. If

$$d \geq 3n + 2L \tag{7}$$

then

$$\Delta^o(n,L) > 0 \quad \text{for } n < n^o \text{ or } L < L^o.$$

Assertions 2 and 3 allow us to formulate a normalized residual plot. We plot a family of curves: $e^o(n,L)$ versus n with L as parameters. As a consequence, the family of curves for $L \geq L^o$ will converge at n^o and then turn "flat." Therefore the structure (n^o,L^o) can easily be located in most cases. A typical plot is given in Figure 1.

Criterion 2. In the normalized residual plot, in a region of (n,L) defined by (7), locate the value n at which a family of L-curves converge and then turn flat. This determines the value n^o. The smallest value in this family of curves is L^o.

Criteria 1 and 2 are called the normalized residual criterion, abbreviated by NRC.

It is easy to see that the above results can be extended to the case of multiple inputs, with the following modification: In Criterion 1 the quantity (K-3n-2L-1) is replaced by [K-(r+2)n-(r+1)L-r], where r is the number of inputs.

The proofs of all results presented and a set of computer simulations are given in [12]. The computer simulations seem to indicate that these criteria are very reliable, and better than the existing ones.

It may be important to note that the criteria are established without specific assumptions on the distribution of the noise. Therefore, they can be applied to those problems for which the distribution of the noise is not specified, or not given. This is especially attractive to socio-economic problems since their distribution are almost impossible to find or estimate.

IV. A COMPARISON WITH AKAIKE'S CRITERION

Akaike's information criterion (AIC) [4-6] is another method for the determination of the order of a difference equation, while the noise is assumed to be a normal distribution. When the criterion is modified in order to include the systems considered in this paper, it states that the best estimate n^o and L^o will minimize the quantity

$$AIC(n,L) = \ln \sigma^2(n,L) + \frac{2[(r+2)n+(r+1)L]}{K-n-L}$$

where $\sigma(n,L)$ is the maximum likelihood estimate of innovation variance. Comparing this criterion with Criterion 2, they usually give identical results when the noise is normal. When the noise is not normal, the Akaike's criterion usually gives an order higher by one or two. A specific example is given here to demonstrate this point.

Consider the two-input/one-output system,

$$\begin{aligned} & y(t) + y(t-1) + 0.5y(t-2) \\ = \; & u_1(t) + u_1(t-1) + 1.5u_1(t-2) + 2u_1(t-3) \qquad (8) \\ + \; & u_2(t-1) + u_2(t-2) - 0.5u_2(t-3) + v(t) \end{aligned}$$

The sequence $v(t)$ is generated from a uniform distribution with magnitude ranging ± 1, $u_1(t)$ from a uniform distribution with magnitude ranging ± 2, $u_2(t)$ is from the equation

$$u_2(t) = u_1(t) + 0.1u_1^2(t) + 0.4w(t)$$

where w(t) is from a uniform distribution ranging ± 1. Note that there exists a multicolinearity between u_1 and u_2. The estimation of the order is given by Figure 2. The Akaike's criterion (Figure 2a) gives n = 5, while Criterion 2 (Figure 2b) gives the true order n = 3.

It may be more important to compare the systems estimated by the two methods. Since the orders are different, the best way to compare systems are by their unit-step responses. These are given in Figure 3. Curve A is the unit-step response from the actual system, Curve B from the one obtained by Criterion 2, and Curve C by Akaike's criterion. Figure 3a is the unit-step response due to u_1. It is seen that Curve B is closer to the true step response than Curve C. Figure 2b is the unit-step response due to u_2. It is seen that Curve B is very close to Curve A. On the other hand, Curve C is opposite to Curve A. Therefore, when Akaike's model is used for the system to be identified, it will give a false response, not only quantitatively but also qualitatively.

The purpose of the above example is to show that if the Akaike's criterion is applied to those problems for which the distribution of noise is not known, the resulted model may lead to serious erroneous results.

It may be more enlightening to realize that since Akaike's model is of higher order, it fits the output data better than the lower order ones. Therefore, a model which fits the output data better, may not be a better model to represent the system. Unfortunately, this is the usual criterion for the evaluation of the goodness of a model. We propose that a measure of the closeness of unit-step responses should be used as a criterion to determine the goodness of a model, instead of coefficients and/or output residuals.

V. CONCLUSION

The normalized residual criterion is presented. This criterion estimates the structure of a dynamical system as well as the variance of the noise and the noise-to-output signal ratio. This criterion has been applied to the modeling of socio-economic systems [9,10,11]. The extension of this criterion to the case of multiple outputs will be discussed in the Minimal-Dimension-Realization Method (MDR method), which will be presented in the near future.

VI ACKNOWLEDGEMENT

This research is supported in part by the National Science Foundation under

Grant ENG76-00775.

VII. REFERENCES

[1] G. E. P. Box and G. M. Jenkisn, Time Series Analysis, Forecasting and Control, Holden Day, 1970.

[2] C. M. Woodside, "Estimation of the order of linear systems," Automatica, Vol. 7, pp. 727-733, 1971.

[3] E. Tse and H. Weinert, "Structure determination and parameter identification for multivariable stochastic linear systems," IEEE Trans. on Auto. Control, Vol. AC-20, pp. 603-613, 1975.

[4] A. Akaike, "Autoregressive model fitting for control," ANN. Inst. Statistics Math., Vol. 23, pp. 163-180, 1971.

[5] A. Akaike, "Information theory and an extension of the maximum likelihood principle," Proc. 2nd Int. Symp. Information Theory, pp. 267-281, 1972.

[6] A. Akaike, "Use of an information theoretic quantity for statistical model identification," Proc. 5th Hawaii Int. Conf. System Sciences, pp. 249-250, 1972.

[7] E. Parzen, "Some recent advances in time series modeling," IEEE Trans. on Auto. Control, Vol. AC-19, pp. 723-730, 1974.

[8] T. O. Lewis and P. L. Odell, Estimation in Linear Models, Prentice-Hall, 1971.

[9] R. Liu, Y. C. Chang, K. S. Kim and L. C. Suen, "An application of MDR method to modeling of gasoline consumption of the United States," Large Scale Dynamical Systems, Point Lobos Press, pp. 113-125, 1976.

[10] L. H. Lin and L. C. Suen, "An investigation of the dynamical relation between the monetary base and M2 by the MDR method," EE Memo. 7602A, University of Notre Dame, Notre Dame, Indiana, June 1976.

[11] R. Liu and L. C. Suen, "An MDR model of the United States Economy," Conference of IEEE System, Man and Cybernetics, (to be presented).

[12] L. C. Suen and R. Liu, "A Normalized Residual Criterion for the Determination of Structure of Linear Systems," EE Memo. 7604, University of Notre Dame, Notre Dame, Indiana, June 1976.

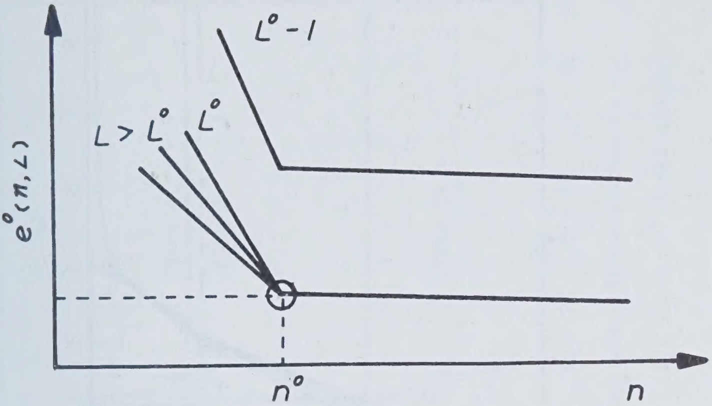
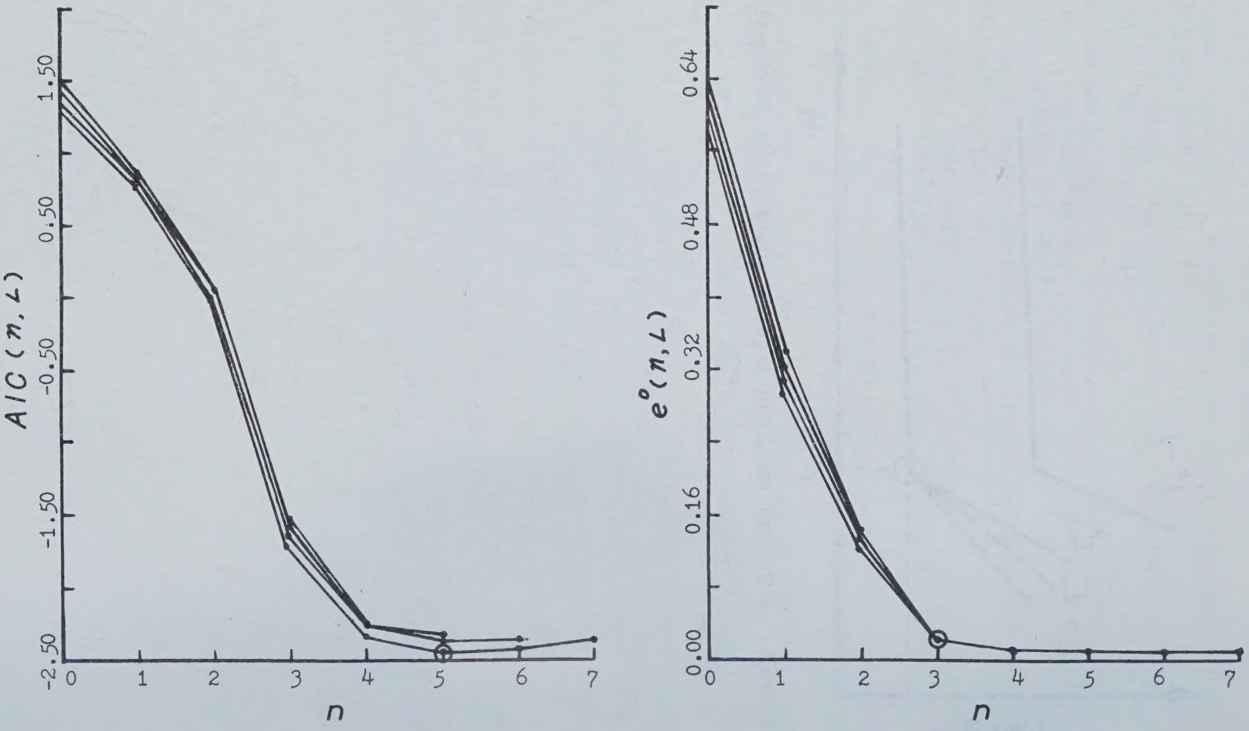


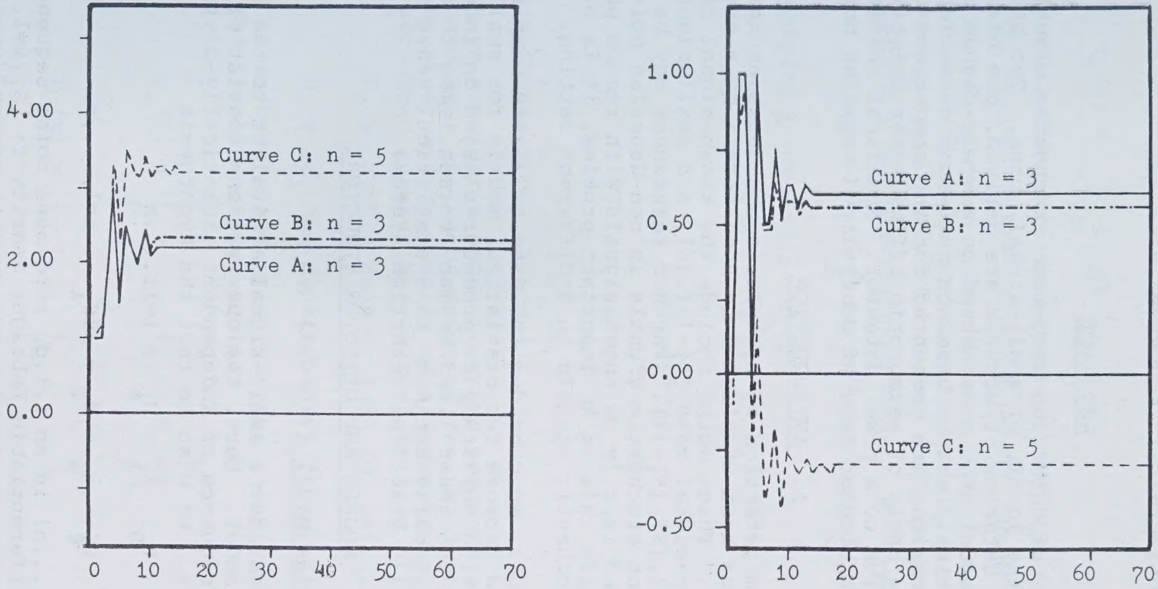
Figure 1. A Typical Family of $e^o(n,L)$ Curves in the Stochastic Case.



(a)

(b)

Figure 2. The Estimation of the Order of Example (12) by (a) AIC and (b) NRC.



(a) Unit-Step Responses due to u_1

(b) Unit-Step Responses due to u_2

Figure 3. The Comparison of the Unit-Step Responses of Example (12): Curve A (Actual System), Curve B (NRC), Curve C (AIC).

LOCAL DETECTION OF STOCHASTIC SIGNALS IN ADDITIVE NOISE

H.V. POOR AND J.B. THOMAS
Department of Electrical Engineering
Princeton University, Princeton, N.J.

ABSTRACT

The problem of detecting zero-mean stochastic signals is considered for the local, or small-signal, case. Two general locally-optimum detector structures are derived, one based on an additive-noise model, the other based on a scale-change model. A third (sub-optimal) scheme based on null-zone detection is also proposed. Expressions are presented for the comparison of these detectors on the basis of asymptotic efficiencies relative to each other and relative to a power detector. Numerical results are given for the particular case of generalized Gaussian noise.

I. INTRODUCTION

The optimum detection of signals in non-Gaussian noise is a topic which has been considered in some detail by a number of authors [1]-[7]. These works include the known-signal case [4] -[7], stochastic-signal case [1]-[3],[6], and small-signal approximations [1],[3],[5]-[7]. (Further references may be found in [7].) Work on stochastic signals in non-Gaussian noise has been concentrated largely on known signals with random phases or amplitudes. While this is an important problem, it is of interest to study stochastic signals in a different setting.

In this paper, we consider the case of a zero-mean stochastic signal and propose two statistical models for small-signal analysis. Limiting our study to procedures based on independent samples, we use the generalized Neyman-Pearson lemma to derive tests which are locally-optimum in a statistical sense and which lead to physically practical detection systems.

II. MODELS AND DETECTOR STRUCTURE

A. Additive-Noise Model (θ-model)

First we consider a small-signal equivalent to the standard additive-noise model. Here, the observations consist of $\{x_i;i=1,\ldots,n\}\equiv\underline{x}$, a sequence of independent, identically-distributed (i.i.d.) samples. We wish to test the hypothesis

$$H_0:\ X_i = N_i \quad ; i=1,\ldots,n$$

versus (1)

$$H_\theta:\ X_i = N_i + \sqrt{\theta}S_i \quad ; i=1,\ldots,n,$$

where $\{N_i;i=1,\ldots,n\}$ is an i.i.d. zero-mean noise sequence with common twice-differentiable Lebesgue density f; $\{S_i;i=1,\ldots,n\}$ is an i.i.d. zero-mean signal sequence with common (possibly unknown) distribution G; and θ is a positive number. We assume the N_i are independent of the S_i and, for the local case, take $\theta\to 0^+$.

The locally-optimum test for (1) is

$$\varphi_\theta(\underline{x}) = \begin{cases} 1 & > k \\ a\ , \quad T_\theta(\underline{x}) & = k \\ 0 & < k \end{cases}$$

where

$$T_\theta(\underline{x}) = \frac{\partial}{\partial\theta} L_\theta(\underline{x}) \Big|_{\theta=0} , \tag{2}$$

L_θ is the likelihood ratio between H_0 and H_θ; k and a are chosen so that $E_0\{\varphi_\theta\} \leq \alpha$, a pre-assigned level; and we choose H_θ with probability $\varphi_\theta(\underline{x})$ when $\underline{x}$ is observed. The generalized Neyman-Pearson lemma [8] asserts that φ_θ maximizes the slope of the power curve, $\partial E_\theta\{\varphi\}/\partial\theta$, at $\theta=0$ among all randomized tests φ for which $E_0\{\varphi\} \leq \alpha$, the pre-assigned level. We have

$$L_\theta(\underline{x}) = \prod_{i=1}^{n} \left[\int_{-\infty}^{\infty} f(x_i - \sqrt{\theta}s)\, dG(s)/f(x_i) \right]$$

and, applying L'Hopital's rule,

$$T_\theta(\underline{x}) = (\overline{s^2}/2) \sum_{i=1}^{n} g_\theta(x_i)$$

where

$$g_\theta(x) \overset{D}{=} f''(x)/f(x) \tag{3}$$

and

$$\overline{s^2} = \int_{-\infty}^{\infty} s^2 dG(s)$$

B. Scale-Change Model (σ-model)

Now we consider a model in which presence of the signal is represented by a change of scale in the noise-only statistics. Again our observations consist of a sequence $\{x_i; i=1,\ldots,n\}$ of i.i.d. zero-mean samples and we wish to test

$$H_0 : X_i \text{ has density } f(x) \text{ for } i=1,\ldots,n ,$$

versus (4)

$$H_\sigma : X_i \text{ has density } f(x/\sigma)/\sigma \text{ for } i=1,\ldots,n ,$$

where f is as before and $\sigma > 1$. For the local case we consider $\sigma \to 1^+$. The locally-optimum test is

$$\varphi_\sigma(\underline{x}) = \begin{cases} 1 & > k' \\ a'\ , \quad T_\sigma(\underline{x}) & = k' \\ 0 & < k' \end{cases}$$

where

$$T_\sigma(\underline{x}) = \frac{\partial}{\partial\sigma} L_\sigma(\underline{x}) \Big|_{\sigma=1} \tag{5}$$

and L_σ is the likelihood ratio between H_σ and H_0. [Note that H_0 in (4) is identical to H_0 in (1).] We have

$$L_\sigma(\underline{x}) = \prod_{i=1}^{n} [f(x_i/\sigma)/\sigma f(x_i)]$$

from which

$$T_\sigma(\underline{x}) = \sum_{i=1}^{n} g_\sigma(x_i)$$

where (6)

$$g_\sigma(x) \overset{D}{=} -(1 + xf'(x)/f(x))$$

Note that detectors φ_θ and φ_σ can both be represented in the form of Fig. 1.

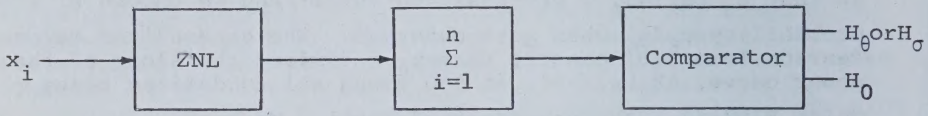


Fig. 1: Locally-Optimum Detector Structure

The samples are passed through a zero-memory nonlinearity (ZNL), summed and compared to a threshold. The two detectors differ only in the shape of the ZNL (g_θ for φ_θ; g_σ for φ_σ).

C. Performance Comparison of φ_θ and φ_σ

Asymptotic relative efficiency (ARE) is frequently used to measure the relative performance of two detectors under small-signal large-sample-size conditions. Recall that, for the situation in (1) the asymptotic efficiency of a detector φ_1 relative to another detector φ_2 is defined as [9]

$$\mathrm{ARE}^{\theta}_{\varphi_1,\varphi_2} = \lim_{\substack{n\to\infty \\ \theta\to 0}} e(\alpha,\theta,n) \tag{7}$$

where $e(\alpha,\theta,n)$ is the relative number of samples detector φ_2 requires to achieve the same power that φ_1 achieves for sample size n when both φ_1 and φ_2 are operating at level α and the signal strength is θ. The parameter θ approaches 0 as $n\to\infty$ in a way so that the power converges to a number $\beta\in(\alpha,1)$. For many important cases the ARE is independent of α. For the scale-change model (4) the ARE is similarly defined with θ replaced by σ and the limit $\theta\to 0$ replaced by $\sigma\to 1$.

Of course, the relative usefulness of φ_θ or φ_σ depends on the relative validity of the corresponding model. However, comparisons on the basis of asymptotic efficiencies will give some indication of how φ_θ and φ_σ perform for a particular model. Under certain mild restrictions the relative efficiencies under the two models can be computed using the method of efficacies [10] to give:

$$\mathrm{ARE}^{\theta}_{\varphi_\theta,\varphi_\sigma} = \mathrm{Var}_0(g_\theta)\mathrm{Var}_0(g_\sigma)/\mathrm{Cov}_0^2(g_\theta,g_\sigma) = \mathrm{ARE}^{\sigma}_{\varphi_\sigma,\varphi_\theta} \tag{8}$$

where the superscripts 'θ' and 'σ' indicate which model is being considered, the subscript '0' on the variance and covariance denotes quantities computed under the null hypothesis H_0 (the same for both models), and g_σ and g_θ are obtained from (3) and (6).

Another performance criterion is the asymptotic efficiency of φ_θ or φ_σ relative to the power detector φ_χ given by

$$\varphi_\chi(\underline{x}) = \begin{cases} 1 & > k'' \\ a'' \;, \quad \sum_{i=1}^{n} x_i^2 & = k'' \\ 0 & < k'' \end{cases} \tag{9}$$

For unit-variance noise it can be shown that the relative efficiencies can be computed from the following equations:

$$\mathrm{ARE}^{\sigma}_{\varphi_\sigma,\varphi_\chi} = (\mu_4-1)\ \mathrm{Var}_0(g_\sigma)/4 \tag{10}$$

$$\mathrm{ARE}^{\theta}_{\varphi_\theta,\varphi_\chi} = (\mu_4-1)\ \mathrm{Var}_0(g_\theta)/4 \tag{11}$$

$$\mathrm{ARE}^{\sigma}_{\varphi_\theta,\varphi_\chi} = (\mu_4-1)\mathrm{Cov}_0^2\ (g_\sigma,g_\theta)/4\mathrm{Var}_0(g_\theta) \tag{12}$$

$$\mathrm{ARE}^{\theta}_{\varphi_\sigma,\varphi_\chi} = (\mu_4-1)\mathrm{Cov}_0^2\ (g_\sigma,g_\theta)/4\mathrm{Var}_0(g_\sigma) \tag{13}$$

where

$$\mu_4 = \int_{-\infty}^{\infty} x^4 f(x)\,dx$$

III. EXAMPLE-GENERALIZED GAUSSIAN NOISE

To illustrate the results of Section II, we consider the case where the noise density f is a member of the generalized Gaussian class. This class is defined in terms of its density functions by

$$f(x|c) = (c\eta(c)/2\Gamma(1/c))\exp\{-(\eta(c)|x|^c)\} \;;\quad x\in(-\infty,\infty)$$

where

$$\eta(c) \overset{D}{=} [\Gamma(3/c)/\Gamma(1/c)]^{1/2}$$

Γ is the gamma function, and c is a positive parameter controlling the decay rate of the tails. For $c>0$, $f(x|c)$ has zero-mean and unit variance, making this class a convenient one for studying the effects of density shape on detector structure and performance. Note that c=2 gives the standard Gaussian distribution.

The locally-optimum detection nonlinearities g_θ and g_σ can be calculated from(3) and (6). We have

$$g_\theta(x|c) = c^2[\eta(c)]^{2c}|x|^{2(c-1)} - c(c-1)\ [\eta(c)]^c|x|^{c-2} \tag{14}$$

and

$$g_\sigma(x|c) = c[\eta(c)]^c|x|^c-1 \tag{15}$$

Figure 2 gives half-plane plots of g_θ and g_σ for various values of

c. Note that, for the standard Gaussian density (c=2), we have

$$g_\theta(x|2) = g_\sigma(x|2) = x^2 - 1$$

so that both models lead to the same detector structure for this case. [This is the power detector of Eq. (9).] However, the two detectors are not generally equivalent for arbitrary c.

The relative performance of φ_θ and φ_σ for the generalized Gaussian case can be evaluated using Eqs. (8) and (10)-(13). To use these equations we need the following quantities:

$$\mathrm{Cov}_0^2(g_\theta, g_\sigma|c) = c^6\eta^4(c)\Gamma^2(2-1/c)/\Gamma^2(1/c) \quad ; \quad c > 1 \tag{16}$$

$$\mathrm{Var}_0(g_\sigma|c) = c \quad ; \quad c > 0 \tag{17}$$

$$\mathrm{Var}_0(g_\theta|c) = c^4\eta^4(c)\Gamma(2-3/c)(1-1/c)(3-4/c)/\Gamma(1/c) \quad ; \quad c > 3/2 \tag{18}$$

$$\mu_4(c) = \Gamma(5/c)/[\Gamma(1/c)\eta^4(c)] \quad ; \quad c > 0 \tag{19}$$

(The ranges of c given with each quantity indicate the range of c for which the ARE expressions in which that quantity appears are valid. For c outside of these ranges the method of efficacies does not apply and no simple method of calculating ARE's is available.) Using these expressions, the efficiencies of Eqs. (8) and (10)-(13) have been computed and are plotted in Fig. 3. Note that, for c near 2, the performance of either detector under the alternate model is near that of the alternate optimum detector.

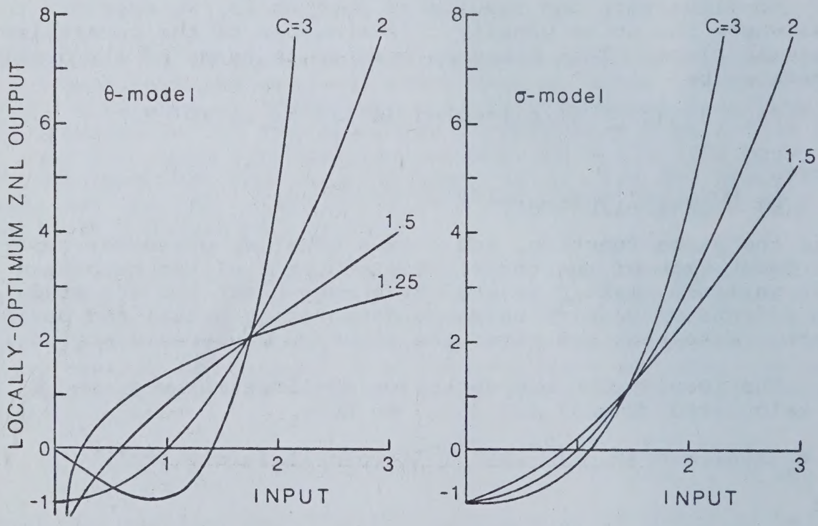


Fig. 2: Optimum ZNL Characteristics (Generalized Gaussian Noise)

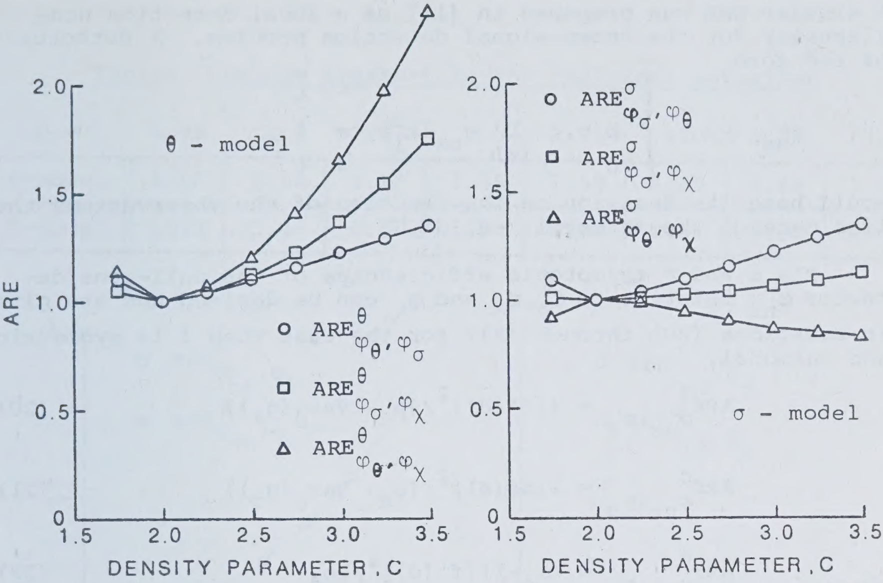


Fig. 3: Asymptotic Performance Comparison of φ_θ, φ_σ, and φ_χ

IV. A SIMPLER DETECTION SCHEME

The ZNL's encountered in the above detector structures are, except for a few special cases, non-standard devices. It may be of interest to use a simple approximation to these devices in order to reduce implementation problems.

One of the simplest possible structures is a detector of the form of Fig. 1 based on the null-zone ZNL given by:

$$g_{nz}(x) = \begin{cases} 1 & , \text{ if } |x| \geq d \\ 0 & , \text{ if } |x| < d \end{cases}$$

i.e., g_{nz} has the following characteristic:

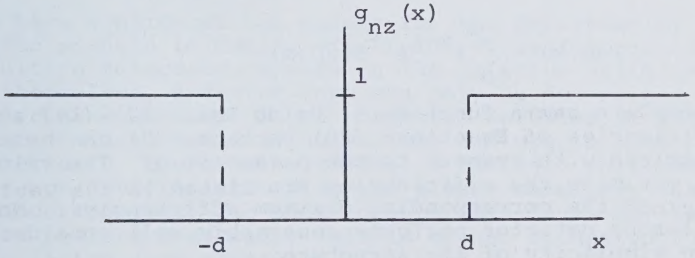


Fig. 4: Null-Zone ZNL

A similar ZNL was proposed in [11] as a local detection nonlinearity for the known-signal detection problem. A detector of the form

$$\varphi_{nz}(\underline{x}) = \begin{cases} 1 & \\ b & , \sum_{i=1}^{n} g_{nz}(x_i) \\ 0 & \end{cases} \begin{matrix} > \ell \\ = \ell \\ < \ell \end{matrix}$$

would base its decision on the fraction of the observations that lies outside the interval (-d,+d).

The θ and σ asymptotic efficiencies of the null-zone detector φ_{nz} relative to $\varphi_\sigma, \varphi_\theta$ and φ_χ can be derived and are **given** in Equations (20) through (23) for the case when f is symmetric and unimodal.

$$\mathrm{ARE}^{\theta}_{\varphi_{nz},\varphi_\theta} = 2[f'(d)]^2/[p_d \cdot \mathrm{Var}_0(g_\theta)] \tag{20}$$

$$\mathrm{ARE}^{\sigma}_{\varphi_{nz},\varphi_\sigma} = 2[df(d)]^2/[p_d \cdot \mathrm{Var}_0(g_\sigma)] \tag{21}$$

$$\mathrm{ARE}^{\theta}_{\varphi_{nz},\varphi_\chi} = (\mu_4 - 1)[f'(d)]^2/2p_d \tag{22}$$

$$\mathrm{ARE}^{\sigma}_{\varphi_{nz},\varphi_\chi} = (\mu_4 - 1)[df(d)]^2/2p_d \tag{23}$$

where

$$p_d \overset{D}{=} [1-F(d)][2F(d)-1]$$

$$F(d) = \int_{-\infty}^{d} f(x)dx$$

and

$$\mu_4 = \int_{-\infty}^{\infty} x^2 f(x)dx$$

For generalized Gaussian densities, $F(d|c)$ is

$$F(d|c) = 1/2 + P(1/c, [\eta(c)d]^c)/2 \tag{24}$$

where

$$P(\alpha,x) = \int_0^x t^{\alpha-1} e^{-t} dt/\Gamma(\alpha)$$

the incomplete gamma function. Using Eqs. (17)-(19) and (24), the efficiencies of Equations (20) through (23) can be computed and maximized with respect to the parameter d. The values of d which maximize the efficiencies are **listed** in the Table and Fig. 5 plots the corresponding maximum efficiencies. Note that the null-zone detector performs reasonably well considering the relative simplicity of its structure.

Table: Optimum Breakpoints for Null-Zone Detection

c =	1.75	2.0	2.25	2.5	2.75	3.0	3.25	3.5
σ-model	1.43	1.48	1.52	1.55	1.58	1.60	1.62	1.63
θ-model	1.12	1.48	1.69	1.83	1.91	1.97	2.01	2.03

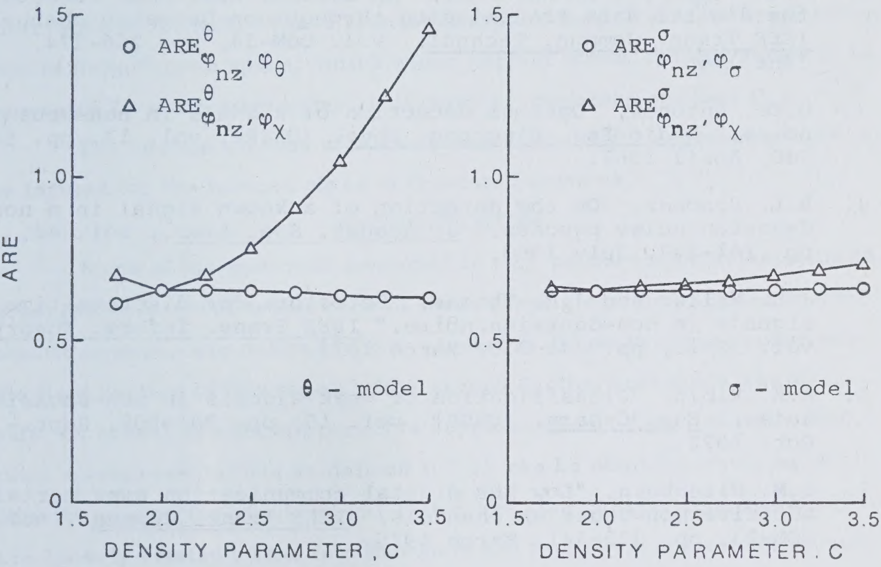


Fig. 5: Asymptotic Efficiencies of φ_{nz} Relative to φ_θ, φ_σ, and φ_χ

V. CONCLUSIONS

We have considered two models for the detection of small stochastic signals in noise. The relative usefulness of the two resulting detectors depends on the relative validity of the models themselves, a factor that can only be determined from the physical characterstics of the channel. We noted in Section III that, for standard Gaussian noise, the two models lead to identical detectors. The results of Fig. 3 indicate that, when the noise is non-Gaussian, choosing the proper model may enhance performance; however, for this (generalized-Gaussian) case, both detectors perform reasonably well in either situation for some large deviation from normality (e.g.,c=3.5).

ACKNOWLEDGEMENTS

This research is supported by the National Science Foundation under Grant-24187 and by the U.S. Army Research Office under Contract DAAG29-75-G-0192.

REFERENCES

1. P. Rudnick, "Likelihood detection of small signals in stationary noise," J. Appl. Phys., vol. 32, pp. 140-143, Feb. 1961.

2. S.S. Rappaport and L. Kurz, "An optimal nonlinear detector for digital data transmission through non-Gaussian channels," IEEE Trans. Commun. Technol., vol. COM-14, pp. 266-274, June 1966.

3. O.Ye. Antonov, "Optimum detection of signals in non-Gaussian noise," Radio Eng. Electron. Phys. (USSR), vol. 12, pp. 541-548, April 1967.

4. R.L. Spooner, "On the detection of a known signal in a non-Gaussian noise process," J. Acoust. Soc. Amer., vol. 44, pp. 141-147, July 1968.

5. J.H. Miller and J.B. Thomas, "Detectors for discrete-time signals in non-Gaussian noise," IEEE Trans. Inform. Theory, vol. IT-18, pp. 241-250, March 1972.

6. A.K. Ribin, "Classification of weak signals in non-Gaussian noise," Eng. Cybern. (USSR), vol. 10, pp. 901-909, Sept.-Oct. 1972.

7. L.M. Nirenberg, "Low SNR digital communication over certain additive non-Gaussian channels," IEEE Trans. Commun., vol. COM-23, pp. 332-341, March 1975.

8. T.S. Ferguson, Mathematical Statistics. New York: Academic, 1967.

9. G.E. Noether, Elements of Nonparametric Statistics. New York: Wiley, 1967.

10. J. Capon, "On the asymptotic efficiency of locally optimum detectors," IRE Trans. Inform. Theory, vol. IT-7, pp. 67-71, April 1961.

11. S.A. Kassam and J.B. Thomas, "The dead-zone limiter: an application of conditional tests in nonparametric detection," J. Acoust. Soc. Amer., to be published late 1976 or early 1977.

PROPERTIES OF CERTAIN CLASSES OF CAUSAL, NONLINEAR, TIME-VARYING STOCHASTIC SYSTEMS.

William L. Root
The University of Michigan, Ann Arbor

ABSTRACT

From the input-output map of a stochastic, nonlinear, time-varying, causal, bounded memory system operating for all time, the input-output maps of truncations of the system corresponding to finite observation intervals are formed. These truncated system maps describe trajectories in a certain Banach space, which under certain conditions are shown to be generated by a one-parameter semigroup of operators of class C_0.

For certain classes of such systems causal ϵ - representations can be formed for the induced class of truncated systems.

INTRODUCTION

Much of the material developed in [2] on the trajectories of finite-observation-interval truncations of time-varying systems applies to stochastic systems are defined below. A sketch of how this goes constitutes the first Section of this paper. The second Section contains a simple result on causal system approximate representation: under certain conditions ϵ -representations as defined in [1] can be obtained which have not only the properties specified in [1] but are also causal. The two Sections are loosely related by the fact that each has to do with one aspect of one kind of theory of system models.

A few concepts and notations are common to both Sections. From [1] we take the definition: a class of systems is $\mathcal{S} = (\mathcal{Y}, f, \mathfrak{X}, \mathcal{U})$ where $\mathcal{Y}$ is a Banach space, $\mathfrak{X}$ and $\mathcal{U}$ are metric spaces and f is a continuous map from the topological product space $\mathfrak{X} \times \mathcal{U}$ into $\mathcal{Y}$. $\mathcal{U}, \mathcal{Y}$ and $\mathfrak{X}$ are, respectively, the input space, output space and system parameter space. Let $\mathcal{F} = \mathcal{F}(\mathcal{U}, \mathcal{Y})$ be the set of all bounded continuous maps from $\mathcal{U}$ into $\mathcal{Y}$ made into a Banach space with the norm: $\| F \| \overset{d}{=} \sup_u \| F(u) \|, u \in \mathcal{U}$, $F \in \mathcal{F}$. If $f(x, \cdot)$ is bounded, and hence a member of $\mathcal{F}$, we write $F(\cdot) = f(x, \cdot)$; let $\mathcal{H}$ be the subset of $\mathcal{F}$ consisting of all such F, and define $g: \mathcal{H} \times \mathcal{U} \to \mathcal{Y}$ by $g(F, u) = F(u)$. Then $\mathcal{S}_0 = (\mathcal{Y}, g, \mathcal{H}, \mathcal{U})$ is a class of systems which is called the natural representation of $\mathcal{S}$. In terms of the natural representation, the individual "systems" become simply the maps F from $\mathcal{U}$ into $\mathcal{Y}$, and with $\mathcal{U}$ and $\mathcal{Y}$ understood we refer simply to "the class of systems $\mathcal{H}$." It is convenient to introduce a notation for the mapping that carries $x \in \mathfrak{X}$ to $F \in \mathcal{H}$. We define the natural mapping ψ: $\mathfrak{X} \to \mathcal{H}$ by $\psi(x) = F$ when $F(\cdot) = f(x, \cdot)$.

We are dealing here entirely with situations for which $\mathcal{U}$ and $\mathcal{Y}$ are spaces of functions (or equivalence classes of functions) of a real variable t, which is interpreted as time. We shall suppose that these functions take values in R^n for some n, unless they are stochastic-valued, in which case their "values" will be n-dimensional random vectors. The operators P_t, projection on the past, and L_t, translation to the left, are defined:

$$[P_t z](s) = z(s) \text{ if } s \leq t, \ = 0 \text{ otherwise.}$$

$$[L_t z](s) = z(s+t).$$

They will apply to all spaces $\mathcal{U}$ and $\mathcal{Y}$, since the assumptions on $\mathcal{U}$ and $\mathcal{Y}$ will require closure under these operations. We also write $R_t = L_{-t}$. A map $F \in \mathcal{F}(\mathcal{U}, \mathcal{Y})$ is causal if

$$P_t F(u) = P_t F(P_t u) \qquad \forall t;$$

it is causal with bounded memory d if

$$[P_{t+T} - P_t] \ F(u) = [P_{t+T} - P_t] \ F([P_{t+T} - P_{t-d}] u) \qquad \forall t, \ \forall T > 0. \tag{1}$$

The subsets of causal, resp., causal-with-bounded-memory-d maps in $\mathcal{F}$ are denoted $\mathcal{F}^o$, resp., $\mathcal{F}^o_d$. These notations agree with those in [2].

TRAJECTORIES OF TRUNCATED STOCHASTIC SYSTEMS

Preliminaries

We say a system is partly stochastic if the outputs are stochastic. Thus a class of systems $\mathcal{S}$ is regarded as partly stochastic if the elements of the Banach space $\mathcal{Y}$ have an interpretation as stochastic quantities. Note that with this definition, any properties of classes of systems that do not depend on a specialization of $\mathcal{Y}$ are valid for partly stochastic systems (the qualifier partly is dropped henceforth, because only partly stochastic systems are of concern here).

In this Section it is to be understood that any class of systems under consideration is in natural representation form, $\mathcal{S}_o = (\mathcal{Y}, g, \mathcal{H}, \mathcal{U})$. It is assumed that $\mathcal{H}$ is a subset (metric subspace) of $\mathcal{F}^o_d\ (\mathcal{U}, \mathcal{Y})$ for some fixed $d > 0$. It is also assumed that $\mathcal{U}$ and $\mathcal{Y}$ are particular kinds of spaces now to be described.

Let T be a fixed positive number and $\|\cdot\|$ denote the ordinary L_2 norm. The linear space $\mathcal{L}^{(2)}_T$ of all equivalence classes of functions, $R \to R^n$, for which the norm

$$\|u\|_T \overset{d}{=} \sup_t (\|(P_{t+T}-P_t)\,u\|),\ t \in R^1 \tag{2}$$

is finite is a Banach space (see [2]). For an arbitrary fixed $T > 0$, let $T' = T+d$. $\mathcal{U}$ is to be a subset (metric subspace) of $\mathcal{L}^{(2)}_{T'}$ which satisfies the following conditions: (1) is shift-invariant, i.e. $u \in \mathcal{U} \Rightarrow L_t u \in \mathcal{U}$, for all t; (2) $\mathcal{U}$ has the projection property (P), i.e. $u \in \mathcal{U} \Rightarrow P_a u \in \mathcal{U}$, $(u-P_b u) \in \mathcal{U}$ and $(P_a-P_b)u \in \mathcal{U}$ for all $a, b \in R$; $\mathcal{U}$ is T-compact, i.e. $(P_a-P_b)\mathcal{U}$ is a compact subset of $L_2[a,b]$ for every finite interval $[a,b]$. See [2] for a discussion of these properties and a proof that given any compact subset $\mathcal{K}$ of $L_2[o,T']$ there is a $\mathcal{U}$ with the above properties such that $(P_{T'}-P_o)\mathcal{U} \supset \mathcal{K}$.

$\mathcal{Y}$ is to be a space $\mathcal{S}\mathcal{L}^{(2)}_T$, which is a stochastic snalogue of $\mathcal{L}^{(2)}_T$. To define $\mathcal{S}\mathcal{L}^{(2)}_T$, consider stochastic processes $\{y(t,\omega), t \in R\}$ where each $y(t,\cdot)$ is an n-dimensional random vector on a complete probability space $(\Omega, \mathcal{A}, P)$. The y are to be measurable processes, i.e. the map $y\colon R \times \Omega \to R^n$ is to be measurable $(\mathcal{L} \times \mathcal{A})$ where $\mathcal{L}$ is the σ-algebra of Lebesgue measurable sets. The y are to be such that

$$E|y(t)|^2 = \int_\Omega |y(t,\omega)|^2\, dP(w) < \infty, \quad t \in R,$$

and for any fixed $T>o$,

$$\sup_{a\in R} \int_a^{a+T} E\,|y(t\,|^2\, dt < \infty.$$

In the above expressions $|y(t,\omega)|$ can be taken to be the Euclidean norm of the n-vector $y(t,\omega)$; also the probability variable is deleted, as in customary, whenever possible. Two processes, y and y' are equivalent if $E\,|y(s) - y'(s)|^2 = 0$ a.e., Lebesgue. The set of all equivalence classes of such processes is a normed linear space with norm

$$\|y\|^s_T \overset{d}{=} \sup_t (E\,\|\,[P_{t+T}-P_t]\ y(\cdot,\omega)\,\|^2)^{1/2}, \tag{3}$$

where again the norm on the right is the ordinary L_2-norm. This space is $\mathcal{S}\mathcal{L}^{(2)}_T$.

(A) $\mathcal{SL}_T^{(2)}$ is a Banach space.

Proof: The proof is a trivial modification of Prop. I in [2]. |||

It is trivially verifiable that if $T_1 \neq T_2$, $\mathcal{SL}_{T_1}^{(2)}$ and $\mathcal{SL}_{T_2}^{(2)}$ are spaces comprised of the same elements and the norms are equivalent.

Remark: The spaces $\mathcal{U}$ and $\mathcal{Y}$ have been specified rather narrowly to keep this Section brief and simple. It is trivial to generalize to $\mathcal{L}_T^{(p)}$ and $\mathcal{SL}_T^{(p)}$, defined analogously to $\mathcal{L}_T^{(2)}$ and $\mathcal{SL}_T^{(2)}$ with respect to L_q-norms, $1 \leq q < \infty$. More importantly, infinite-memory systems can be handled by introducing $\mathcal{L}_\phi^{(2)}$, which is defined with respect to an arbitrary memory-weighting function ϕ, as in [3]. One can allow stochastic inputs by having $\mathcal{U}$ a subset of $\mathcal{SL}_T^{(2)}$ or $\mathcal{SL}_\phi^{(2)}$. Finally, a more fundamental kind of generalization is possible; see Remark at end of this Section.

Truncated system trajectories

The output record over an interval of length T starting at t is given by Eq. (1). With T fixed, $0 < T < \infty$, define F_t' for all $t \in R$ by

$$F_t'(u) = [P_{t+T} - P_t] F(u), \quad u \in \mathcal{U} \tag{4}$$

Because of Eq. (1), F_t' is a map from $[P_{t+T} - P_{t-d}]\mathcal{U}$ into $[P_{t+T} - P_t]\mathcal{Y}$. Since $\mathcal{U}$ is required to be shift-invariant and $\mathcal{Y}$ is automatically so, elements of $[P_{t+T} - P_{t-d}]\mathcal{U}$ may be identified with those of $\mathcal{U}_T = [P_T - P_{-d}]\mathcal{U}$ and elements of $[P_{t+T} - P_t]\mathcal{Y}$ may be identified with those of $\mathcal{Y}_T = [P_T - P_o]\mathcal{Y}$. Thus F_t' corresponds to a map $F_t: \mathcal{U}_T \to \mathcal{Y}_T$ defined by

$$F_t = [P_T - P_o] \, L_t F R_t \tag{5}$$

$\mathcal{U}_T$ inherits a metric from $\mathcal{U}$ since elements of $\mathcal{U}_T$ correspond to elements of $\mathcal{U}$ by property (P); this means simply that $\mathcal{U}_T$ has the metric given by the L_2-norm on $[-d, T]$. $\mathcal{U}_T$ is a compact subset of $L_2[-d,T]$. $\mathcal{Y}_T$ can be identified with the L_2-space formed with respect to the product of the probability measure and Lebesgue measure on $[0, T]$. It follows easily that $F_t \in \mathcal{F}_d^o(\mathcal{U}_T, \mathcal{Y}_T)$.

As t varies F_t describes a trajectory in the space $\mathcal{F}_d^o(\mathcal{U}_T, \mathcal{Y}_T)$. For fixed T each F has such a trajectory, and these trajectories characterize the evolutionary changes of the system with time. This is all exactly as in [2]. The only difference here is in the space $\mathcal{Y}_T$, and once certain simple properties are established for the stochastic $\mathcal{Y}_T$ there remain no essential differences as far as a discussion of the trajectories is concerned.

The first property needed is continuity of the shift operation on $\mathscr{Y}_T$ as a function of the shift, in the following sense.

(B) Let $y \in \mathscr{SL}_T^{(2)}$ and vanish outside $[0, T]$, then $E \,||\, L_h y - y \,||^2 \to 0$ as $h \to 0$, where again the norm is the ordinary L_2-norm.

Proof: Let $h_o > o$ be fixed and restrict h, $|h| \leq h_o$. Since y vanishes outside $[0, T]$,

$$E\,||L_h y - y||^2 = \int_\Omega \int_{-h_o}^{T+h_o} |y(t+h, \omega) - y(t, \omega)|^2 dt\, dP(\omega).$$

$(E\,||\cdot||^2)^{1/2}$ is the L_2-norm with respect to the measure $(dt\, dP)$; for convenience we denote it $[\![\cdot]\!]$. Given $\epsilon > 0$, let $\tilde{y}(t, \omega)$ be the process equal to $\min[y(t, \omega), b]$, where $b > 0$ is large enough that $[\![y - \tilde{y}]\!] \leq \epsilon$. Then,

$$[\![L_h y - y]\!] \leq [\![L_h y - L_h \tilde{y}]\!] + [\![L_h \tilde{y} - \tilde{y}]\!] + [\![\tilde{y} - y]\!]$$

where the first and last terms are $\leq \epsilon$.

By the conditions on y, the definition of $\tilde{y}$ and the Fubini theorem,

$$\phi(h, \omega) \overset{d}{=} \int_{-h_o}^{T+h_o} |\tilde{y}(t+h) - \tilde{y}(t)|^2 dt$$

is defined for all ω except possibly for $\omega \in \Lambda$ where $P(\Lambda) = 0$, and for all h. For $\omega \in \Lambda$, $|\phi(h, \omega)|$ is less than a fixed constant; also $\phi(h, \omega) \to 0$ as $h \to 0$ by the classical theorem for shifts of square-integrable functions. Then,

$$[\![L_h y - y]\!]^2 = \int_\Omega \phi(h, \omega)\, dP(\omega) \to 0 \text{ as } h \to 0$$

by the bounded convergence theorem. Thus for h sufficiently small $[\![L_h y - y]\!] \leq 3\epsilon$. |||

The second property needed is that for $y \in \mathscr{SL}_T^{(2)}$, $[\![(P_{t+h} - P_t) y]\!] \to 0$ as $h \to 0$. This is immediately obvious. We now have:

(C) With $\mathscr{U}$ and $\mathscr{Y}$ as given in this Section and $F \in \mathscr{F}_d^o(\mathscr{U}, \mathscr{Y})$, the trajectory F_t is a continuous map from R into $\mathscr{F}(\mathscr{U}_T, \mathscr{Y}_T)$.

Proof: Using the two properties mentioned for $\mathscr{SL}_T^{(2)}$ one can retrace step-by-step the proof of Proposition 12 in [2] (there is a small error in that proof which can be rectified easily). This gives the result. |||

Most of what follows Proposition 12 in [2] holds also with the stochastic output space $\mathscr{SL}_T^{(2)}$. We mention only the generation of trajectories F_t

under certain conditions by a strongly continuous semigroup of linear operators. The conditions are stated in the form given in [3] in a discussion of deterministic systems with infinite memory.

Consider a class of systems $\mathcal{H} \subset \mathcal{F}^{o}_{d}(\mathcal{U}, \mathcal{Y})$. Let $\mathcal{M} = V(\mathcal{H})$, the linear span of $\mathcal{H}$. $\mathcal{F}^{o}_{d}$ is a closed linear subspace of $\mathcal{F}$ with $\mathcal{U}$ and $\mathcal{Y}$ as specified, so $\mathcal{M} \subset \mathcal{F}^{o}_{d}$. Let π_t be defined by $\pi_t F = F_t$; π_t is a linear transformation from $\mathcal{F}^{o}_{d}(\mathcal{U}, \mathcal{Y})$ into $\mathcal{F}^{o}_{d}(\mathcal{U}_T, \mathcal{Y}_T)$ with linear operator norm ≤ 1. Put $\mathcal{M}_t = \pi_t \mathcal{M}$. Consider the following condition:

Condition (sg): If $\pi_t F = \pi_s G$ for some $t, s \geq 0$, then $\pi_{t+a} F = \pi_{s+a} G$ for all $a \geq 0$.

For $H \in \mathcal{M}_t$, write $\pi_t^{-1}(H) = \{F \in \mathcal{M} : \pi_t F = H\}$. If condition (sg) is satisfied one can define $\theta(t, s): \mathcal{M}_t \rightarrow \mathcal{M}_s$, $0 \leq t \leq s$, by

$$\theta(t, s) H = \pi_s \circ \pi_t^{-1} H$$

θ is linear. It follows easily, again using condition (sg), that one can further define an operator $\theta(\tau)$ by

$$\theta(\tau) H = \theta(t, s) H, \quad s = t+\tau, \quad \tau \geq 0, \quad t \geq 0$$

for all H such that for some t, $H \in \mathcal{M}_t$. $\theta(\tau)$ is well-defined; its original domain for any τ is $\cup_{t \geq 0} \mathcal{M}_t$. This can be extended by linearity to $\mathcal{N}_+ \overset{d}{=} V(\cup_{t \geq 0} \mathcal{M}_t)$.

The family $\{\theta(\tau), \tau \geq 0\}$ is a one-parameter semigroup of linear transformations on $\mathcal{N}_+$. If the $\theta(\tau)$ are bounded linear operators they may be extended by continuity to $\overline{\mathcal{N}}_+$, the closure of $\mathcal{N}_+$ in $\mathcal{F}^{o}_{d}(\mathcal{U}_T, \mathcal{Y}_T)$. This is all as in the deterministic cases of [2] and [3] , and again as a corollary of the result on continuity of trajectories, one has:

(D) With fixed $T > 0$ let $\mathcal{H} \subset \mathcal{F}^{o}_{d}(\mathcal{U}, \mathcal{Y})$ satisfy the condition (sg), and let the $\theta(\tau)$, $\tau \geq 0$, be bounded operators. Then $\{\theta(\tau), \tau \geq 0\}$ is a one-parameter operator semigroup of class C_o (cf. Proposition 20 in [2]).

We also have:

(E) If $\mathcal{H}$ satisfies the hypotheses of (D) with $T' > 0$, then it also satisfies the hypotheses and hence the conclusion with any $T > T'$.

Remark: It may appear to the reader that in some sense the particular spaces used for $\mathcal{Y}$, $\mathcal{Y}_T$, and $\mathcal{U}$ and $\mathcal{U}_T$ should not matter very much as far as the characterizations of the trajectories are concerned. In [4] an abstract analogue of the development in [2] is given that does not involve particular spaces, but only families of spaces with certain characteristics. This allows isolating the properties of the spaces involved that do lead to results like (C), (D) and (E).

CAUSAL ϵ -REPRESENTATIONS

Preliminaries

Given a system $\mathcal{S}_0 = (\mathcal{Y}, g, \mathcal{H}, \mathcal{U})$ in natural representation form, an ϵ -representation of $\mathcal{S}_0$ is a pair $(\mathcal{S}_1, \phi)$, where $\mathcal{S}_1$ is a system $(\mathcal{Y}, f_1, \mathcal{X}_1, \mathcal{U})$, ϕ is a map from $\mathcal{H}$ into $\mathcal{X}_1$ and the condition $\| F - \psi_1 o \phi (F) \| \leq \epsilon$ is satisfied for all $F \in \mathcal{H}$ (see [1]). The norm is the norm of $\mathcal{F}(\mathcal{U}, \mathcal{Y})$, ψ_1 is the natural mapping for $\mathcal{S}_1$. The concept of causality is not necessarily relevant to ϵ -representations, but if $\mathcal{U}$ and $\mathcal{Y}$ are spaces of functions of time as in the previous Section and $\mathcal{H} \subset \mathcal{F}^o(\mathcal{U}, \mathcal{Y})$, it may be. Causality of ϵ -representations is not considered in [1] . However, it is shown in [1] that if $\mathcal{H}$ and $\mathcal{U}$ are compact there always exists an ϵ -representation, for arbitrary $\epsilon > 0$, with a number of desirable properties, viz.: (1) $\mathcal{X}_1$ is a subset of some R^n; (2) ϕ and ψ_1 are continuous; (3) ϕ depends only on the functions F, $F \in \mathcal{H}$, as restricted to finitely many elements of $\mathcal{U}$; (4) ψ_1 is a restriction of a linear map. In addition, (5), if $\mathcal{Y}$ is a Hilbert space, ϕ is a restriction of a linear map. The first three properties listed are essential in order than an ϵ -representation provide a suitable model for identification; the linearity is useful. If one considers the classes of truncated systems defined in the previous Section it will be noticed that $\mathcal{U}_T$ is compact and $\mathcal{Y}_T$ is a Hilbert space. Hence with the additional assumption that $\mathcal{H}$ be compact all the conditions are satisfied to guarantee that an ϵ -representation exists with all the properties listed.

There is nothing unique about ϵ -representations, but at least the construction given in the proof of the existence theorem in [1] will not yield causal representing systems even when the $F \in \mathcal{H}$ are causal (each representing system is basically an interpolation among entire records of input-output data). However, we shall see that in a special case, any family of ϵ -representations for arbitrary $\epsilon > 0$ can be converted into a family of causal ϵ -representations, for arbitrary $\epsilon > 0$, at the cost of increased complexity.

The notion of strict causality is used by Porter, Clark and

DeSantis in [5] in a study of causal polynomial approximation to causal functions on a Hilbert space. The concept is also useful here; it is for uniformly strictly causal systems in an appropriate setting that we show the existence of causal ϵ -representations.

Existence of causal ϵ -representations

For everything that follows the class of systems is $\mathcal{S}_o = (\mathcal{Y}_T, g, \mathcal{H}, \mathcal{U}_T)$, where $\mathcal{Y}_T$ is an L_p-space on $[0, T], 1 \leq p < \infty$; $\mathcal{U}_T = (P_T - P_{-d})\mathcal{U}$ with $\mathcal{U}$ shift-invariant, T-compact, and satisfying property (P), and $\mathcal{H}$ is a compact subset of $\mathcal{F}^o(\mathcal{U}_T, \mathcal{Y}_T)$. With the assumption on $\mathcal{H}$ added, the truncated systems of Section 1 satisfy these conditions, as do their deterministic counterparts in [2] .

On the interval $[0, T]$ consider sets of points $\{t_o, t_1, \ldots, t_N\}$ where $t_i = (i/N)T$. When N is temporarily fixed we write $P_i = P_{t_i}$. Given a mapping $F \in \mathcal{F}^o(\mathcal{U}_T, \mathcal{Y}_T)$, define $F^{N,i}$ and F^N by

$$F^{N,i} = (P_i - P_{i-1})\, F\, P_{i-1}$$

$$F^N = \sum_{i=1}^{N} F^{N,i}. \tag{6}$$

$F \in \mathcal{F}^o(\mathcal{U}_T, \mathcal{Y}_T)$ is strictly causal if $\| F^N - F \| \to 0$ as $N \to \infty$, where the norm is that of $\mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$; $\mathcal{H} \subset \mathcal{F}^o(\mathcal{U}_T, \mathcal{Y}_T)$ is uniformly strictly causal if given $\epsilon > 0$ there is $N(\epsilon)$ such that $\| F^N - F \| \leq \epsilon$ for all $N \geq N(\epsilon)$, for all $F \in$.

(F) Let $\mathcal{S}_o$ satisfy the conditions stated above and in addition let $\mathcal{H}$ be a uniformly strictly causal subset of $\mathcal{F}^o(\mathcal{U}_T, \mathcal{Y}_T)$. Then, given any $\epsilon > 0$ there is an ϵ -representation $(\mathcal{S}_1, \phi)$ which has the properties (1), . . . , (4) and which also has the property that $\psi_1(x)$, $x \in \mathcal{X}_1$, is a causal map from $\mathcal{U}$ into $\mathcal{Y}$. If $\mathcal{Y}_T$ is an L_2 space then ϕ is a restriction of a linear map.

Proof: Fix $\epsilon > 0$ and choose N large enough so that $\| F - F^N \| \leq \epsilon/2$ for all $F \in \mathcal{H}$. For each i we consider new spaces $P_{i-1}\mathcal{U}_T$, and $\Delta_i \mathcal{Y}_T = (P_i - P_{i-1})\mathcal{Y}_T$, and the set $\mathcal{H}^{N,i}$ of maps $F^{N,i}$ regarded as maps from $P_{i-1}\mathcal{U}_T$ into $\Delta_i \mathcal{Y}_T$. The norms of elements of $P_{i-1}\mathcal{U}_T$ and $\Delta_i \mathcal{Y}_T$ are the norms of the corresponding elements of $\mathcal{U}_T$ and $\mathcal{Y}_T$, respectively.

The following assertions can be easily verified: $P_{i-1}\mathcal{U}_T$ is compact; each $F^{N,i}$ is bounded and continuous, and $\mathcal{H}^{N,i}$ is compact. Hence, each class of systems $\mathcal{S}_o^i = (\Delta_i \mathcal{Y}_T, g, \mathcal{H}^{N,i}, P_{i-1}\mathcal{U}_T)$ has an $(\epsilon/2N)$-representation $(\mathcal{S}_1^i, \phi^i)$ with properties (1), ..., (4), by Prop. 7 [1]. Also each $F^{N,i}$ is obviously causal. The notation for the representing class of systems is given by $\mathcal{S}_1^i = (\Delta_i \mathcal{Y}_T, f^i, \mathcal{X}^i, P_{i-1}\mathcal{Y}_T)$. $\mathcal{X}^i$ is contained in R^{n_i}, where n_i is some positive integer, and $\phi^i(F^{N,i}) = x \in \mathcal{X}^i$.

We now can obtain a causal $(\epsilon/2)$-representation for $\mathcal{S}_{oN} = (\mathcal{Y}_T, g, \mathcal{H}^N, \mathcal{U}_T)$, where $\mathcal{H}^N \subset \mathcal{F}^o(\mathcal{U}_T, \mathcal{Y}_T)$ is the set of all F^N for $F \in \mathcal{H}$. Let $\mathcal{X}_1 = \sum_{i=1}^{N} \mathcal{X}^i$; $\phi_1(F^N) = (\phi^1(F^{N,1}), \dots, \phi^N(F^{N,N}))$, and

$$f_1(x,u) = \sum_{i=1}^{N} (P_i - P_{i-1}) f^i(x^i, u^i) = \sum_{i=1}^{N} f^i(x^i, u^i)$$

where $x^i \in \mathcal{X}^i$, $u^i = P_{i-1}u$, and elements of $(P_i - P_{i-1})\mathcal{Y}_T$ are identified with the corresponding elements of $\mathcal{Y}_T$. Then $(\mathcal{S}_1, \phi_1)$, $\mathcal{S}_1 = (\mathcal{Y}_T, f_1, \mathcal{X}_1, \mathcal{U}_T)$, is such an $(\epsilon/2)$-representation, since

$$||F^N(u) - [\psi_1 \circ \phi_1(F^N)]\,(u)|| = ||F^N(u) - f_1(\phi_1(F^N), u)|| = ||\sum_{i=1}^{N} F^{N,i}(u) - \sum_{i=1}^{N} f^i(\phi^i(F^{N,i}), u|| \le \sum_{i=1}^{N} ||F^{N,i}(u) - f^i(\phi^i(F^{N,i}), u)|| \le \sum_{i=1}^{N} \epsilon/2N = \epsilon/2.$$

for all $u \in \mathcal{U}_T$. The approximating maps $H^N(\cdot) = f_1(\phi_1(F^N), \cdot)$ are obviously causal. ϕ_1 and ψ_1 are continuous since the ϕ^i and ψ^i are; ψ_1 is a restriction of a linear map since each ψ^i is. $\phi_1(F^N)$ is determined by the aggregate of all the inputs required to determine the $\phi^i(F^{N,i})$, which is finite in number.

It now follows that $||F - \psi_1 \circ \phi_1(F^N)|| \le \epsilon$ for all $F \in \mathcal{H}$. But $\phi_1(F^N)$ can be replaced by $\phi(F)$ where ϕ is defined by

$$\phi(F) = (\phi^1([P_1 - P_o]\, F\, P_o), \dots, \phi^N([P_N - P_{N-1}]\, F\, P_{N-1})),$$

and where ϕ retains the properties (2) and (3). This proves the main assertion. The linearity of ϕ if $\mathcal{Y}_T$ is an L_2-space follows from the linearity of the ϕ^i in this case. |||

Each $F \in \mathcal{H}$ can be strictly causal without $\mathcal{H}$ being uniformly strictly causal. There is space here only to state a generic example to show that a reasonably wide variety of classes of systems are uniformly strictly causal.

Let $\mathcal{U}_T$ be the unit ball of $L_2[-d, T]$ and $\mathcal{Y}_T = L_2[0, T]$. Let $\mathcal{H}$ consist of maps $F \in \mathcal{F}_d^o(\mathcal{U}_T, \mathcal{Y}_T)$ defined by

$$[\,F(u)]\,(t) = \sum_{k=1}^{K} \int_{-d}^{t} \cdots \int_{-d}^{t} h_k(t, s_1, \dots, s_k)\, u\,(s_1) \dots u(s_k)\, ds_1 \dots ds_k \qquad 0 \leq t \leq T, \tag{7}$$

where K is an arbitrary positive integer, and where the kernels h_k satisfy the following conditions uniformly for all $F \in$ and $k = 1, \dots, K$:

(1) $\int_0^T \int_{-d}^T \cdots \int_{-d}^T |h_k(t, s_1, \dots, s_k)|^2 ds_1 \dots ds_k dt \leq B < \infty$.

(2) Given any $\epsilon > 0$, there is $M(\epsilon) > 0$ such that if

$\tilde{h}_k(t, s_1, \dots, s_k) = h_k(t, s_1, \dots, s_k)$ for $|h_k| \leq M$, $\quad = 0$ for $|h_k| > M$,

then $\int_0^T \int_{-d}^T \cdots \int_{-d}^T |h_k - \tilde{h}_k|^2 ds_1 \dots ds_k\; dt \leq \epsilon$.

Then the system $\mathcal{S}_o = (\mathcal{Y}, g, \mathcal{H}, \mathcal{U})$ is uniformly strictly causal. The proof is not given, but it is straightforward. This example can be generalized to infinite-memory and stochastic kernels.

ACKNOWLEDGEMENT

This research is sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under AFOSR Grant Number 72-2328.

REFERENCES

[1] W. L. Root, "On the modelling of systems for identification. Part I: ϵ - representations of classes of systems," SIAM J. Control 13 (1975), pp. 927-944.

[2] W. L. Root, "On the modelling of systems for identification. Part II: time-varying systems" SIAM J. Control 13 (1975), pp. 945-974.

[3] W. L. Root, "A semigroup representation of linear classes of causal, time-varying systems" Conf. Record Ninth Asilomar Conf. on Circuits Systems and Computers (1975) pp. 238-241.

[4] W. L. Root, "An abstract characterization of time-varying systems" Report in preparation, Univ. of Michigan.

[5] W.A. Porter, T.M. Clark and R.M. DeSantis, "Causality structure and the Weierstrass theorem" J. Math Anal. and Appl. 52 (1975), pp. 351-363.

ON THE REPRESENTATION OF NONLINEAR SYSTEMS WITH GAUSSIAN INPUTS

STEEL T. HUANG
Department of Mathematics
University of Cincinnati
Cincinnati, Ohio 45221

STAMATIS CAMBANIS
Department of Statistics
University of North Carolina
Chapel Hill, North Carolina 27514

ABSTRACT

An arbitrary nonlinear system with input a Gaussian process, which is such that its output process has finite second moments, admits two kinds of representations; the first in terms of a sequence of deterministic kernels and the second in terms of a single stochastic kernel. We consider here the identification of the sequence of deterministic kernels from the input and output processes, the representation of the system output when its input is a sample function of the Gaussian process, and the relationship of the sequence of kernels mentioned above to the Volterra expansion kernels when the system has a Volterra representation.

1. STOCHASTIC AND MULTIPLE WIENER INTEGRALS FOR GAUSSIAN PROCESSES

Let us first introduce our basic notation and terminology. We will consider throughout a zero mean Gaussian process $X = (X_t, t\epsilon T)$ with covariance function $R(t,s)$, defined on a probability space $(\Omega,\mathcal{B},P)$. T will be an interval of the real line. $\mathcal{B}$ is usually taken to be $\mathcal{B}(X)$, the σ-field generated by the process X, or $\overline{\mathcal{B}}(X)$, the completion of $\mathcal{B}(X)$. There are two important Hilbert spaces associated to a Gaussian process. The nonlinear space of X, $L_2(X) = L_2(\Omega,\mathcal{B}(X),P)$, consists of all $\mathcal{B}(X)$-measurable random variables with finite second moment which are called (nonlinear) L_2-functionals of X. The linear space of X, $H(X)$, is the closed subspace of $L_2(X)$ spanned by X_t, $t\epsilon T$, and its elements are called linear L_2-functionals of X.

The first useful notion in the study of the nonlinear space of a Wiener process is the Multiple Wiener Integral. This notion was first introduced by Wiener (1938), who termed it "Polynomial Chaos," and was redefined in a somewhat deeper way by Itô (1951). Itô showed that his multiple integrals of different degree have the important property of being mutually orthogonal and also presented their connection with the celebrated Fourier-Hermite expansion of L_2-functionals of Cameron and Martin (1947). In his important work on nonlinear problems Wiener (1958) reinterpreted the multiple Wiener integrals for a Wiener process in an extremely simple and intuitive way and made some interesting applications.

In [4] multiple Wiener integrals of the following two types are defined for general Gaussian processes:

$$I_p(f_p) = \int\cdots\int f(t_1,\ldots,t_p)dX_{t_1}\ldots dX_{t_p}$$

$$J_p(f_p) = \int\cdots\int f(t_1,\ldots,t_p)X_{t_1}\ldots X_{t_p}\, dt_1\ldots dt_p$$

where $p = 1,2,\ldots$, and we always write $\int$ for $\int_T$. We always assume that $X_{t_0} = 0$ a.s. for some $t_0\epsilon T$ when dealing with integrals I_p, and that X is mean square continuous when dealing with integrals J_p.

The integrands $f(t_1,\dots,t_p)$ belong to appropriate Hilbert spaces of "functions" defined on T^p, which are denoted by $\Lambda_2(\otimes^p R)$ for I_p, and $\lambda_2(\otimes^p R)$ for J_p. For instance $\lambda_2(R)$ is the completion of the set of all functions $f(t)$ on T which are such that the Riemann integral $\mathcal{R}\int\int f(t)f(s)\, R(t,s)dt\, ds$ exists and is finite with respect to the inner product

$$<f,g> = \mathcal{R}\int\int f(t)f(s)\, R(t,s)dt\, ds\ .$$

When $T = [a,b]$, $\lambda_2(R)$ contains all square integrable functions over T, as well as further interesting classes of functions (see [4]), but also "functions" with properties similar to those of delta functions. $\lambda_2(\otimes^p R)$ is defined similarly, and is isomorphic to the tensor product $\otimes^p\lambda_2(R)$ under the natural correspondence.

The multiple Wiener integrals J_p, $p=1,2,\dots$, have the following properties ($f,g\epsilon\lambda_2(\otimes^p R)$ and a,b real numbers):

$$J_p(af+bg) = aJ_p(f) + bJ_p(g)\ ,$$

$$J_p(f) = J_p(\tilde{f})\ ,$$

$$<J_p(f),\ J_p(g)>_{L_2(X)} = p!<\tilde{f},\tilde{g}>_{\lambda_2(\otimes^p R)}\ ,$$

where $\tilde{f}$ is the symmetric version of f,

$$<J_p(f),J_q(q)>_{L_2(x)} = 0 \quad \text{if } p \neq q\ ,$$

$$J_p(\phi_1^{\tilde{\otimes}p_1}\tilde{\otimes}\dots\tilde{\otimes}\,\phi_k^{\tilde{\otimes}p_k}) = H_{p_1}(\int\phi_1(t)X_t dt)\dots H_{p_k}(\int\phi_k(t)X_t dt)\ ,$$

where $\phi_1,\dots,\phi_k$ are orthonormal in $\lambda_2(R)$, $p_1+\dots+p_k = p$, H_p denote the Hermite polynomials, and $\tilde{\otimes}$ denotes symmetric tensor product.

Also every L_2-functional θ of X, $\theta\epsilon L_2(X)$, has an orthogonal development

$$\theta = E(\theta) + \sum_{p=1}^{\infty} J_p(f_p)$$

where $f_p\epsilon\lambda_2(\otimes^p R)$, and if

$$\theta - E(\theta) = \sum_{p=1}^{\infty} J_p(f_p) = \sum_{p=1}^{\infty} J_p(g_p)$$

then $\tilde{f}_p = \tilde{g}_p$, $p=1,2,\dots$.

The second useful notion in the study of the nonlinear space of a Wiener process is the stochastic integral. The stochastic integral was first introduced by Itô (1944) for the Wiener process and later generalized by Meyer for martingales. Every L_2-functional of a Wiener process has a representation as a stochastic integral, where the integrand is adapted to the Wiener process.

In [4] a stochastic integral

$$I(f) = \int f(t)dX_t$$

is defined for $L_2(X)$-valued "functions" $f(t)$ belonging to the Hilbert space $\Lambda_{2;L_2(X)}(R)$, which is defined in a way similar to $\Lambda_2(R)$ and is isomorphic to the tensor product $\Lambda_2(R)\otimes L_2(X)$ under the correspondence $\phi\xi \longleftrightarrow \phi\otimes\xi$, $\phi\in\Lambda_2(R)$, $\xi\in L_2(X)$. The stochastic integral

$$I\colon \Lambda_{2;L_2(X)}(R) \to L_2(X)$$

is an unbounded densely defined closed linear map with range the set of all zero mean random variables in $L_2(X)$ (for further properties and some evaluations of the stochastic integral see [4]). Thus every L_2-functional θ of X, $\theta\in L_2(X)$, admits the representation

$$\theta = E(\theta) + \int f(t)dX_t$$

for some (non-unique) $f\in\Lambda_{2;L_2(X)}(R)$ in the domain of I, and f may be taken to be adapted to X ("adapted to" meaning "measurable with respect to the past of"). Notice that as shown in [4], the stochastic integral becomes norm preserving when restricted to nonanticipatory integrands, but it is not yet known which L_2-functionals admit nonanticipatory stochastic integral representations ("nonanticipatory" meaning "independent of the future increments of X").

2. NONLINEAR SYSTEMS WITH GAUSSIAN INPUTS

Let θ be a nonlinear system with input the Gaussian process $X = \{X_t, t\in T\}$ and output the process $Y = \{Y_t, t\in T\}$ (the parameter sets of the input and output processes could of course be distinct). The only assumption we make on the nonlinear system without further notice is that it is such that the process Y is of second order, i.e. each Y_t is an L_2-functional of X. Then from Section 1, we have two representations of the system θ for the input X. The first is a series representation in terms of multiple Wiener integrals

$$Y_t = E(Y_t) + \sum_{p=1}^{\infty} \int\cdots\int f_p(t;t_1,\ldots,t_p)X_{t_1}\ldots X_{t_p}dt_1\ldots dt_p \tag{1}$$

where $f_p(t;\cdot)\in\Lambda_2(\otimes^p R)$ may (and will from now on) be taken to be symmetric. The second is a stochastic integral representation

$$Y_t = E(Y_t) + \int f(t;s)X_s ds \tag{2}$$

where $f(t;\cdot)\in\Lambda_{2;L_2(X)}(R)$ may be taken to be adapted to X. Thus the system θ can be represented for the input X either by the sequence of deterministic kernels $f_p(t;t_1,\ldots,t_p)$, $p=1,2,\ldots$, as in (1), or by the single stochastic kernel $f(t,s)$ as in (2). It should be emphasized that both, the sequence of deterministic kernels $\{f_p\}$ and the stochastic kernel f, depend not only on the system θ but also on the input Gaussian process X. Thus distinct input processes will always produce distinct stochastic kernels (unless of course the system is linear) in representation (2), and will in general produce distinct sequences of deterministic kernels in representation (1).

From now on we concentrate on the representation (1). Notice that expansion (1) looks very much like a Volterra kernel expansion, with the important difference that the multiple integrals are multiple Wiener rather than Lebesgue integrals; it could thus be considered as a stochastic Volterra kernel expansion. Several system synthesis or identification problems naturally suggest themselves at this point:

(P1) Knowing the input and output processes X and Y, find the kernels f_p, $p=1,2,\ldots$.

(P2) Knowing the kernels f_p, $p=1,2,\ldots$, in the representation of the system for the Gaussian input X, can one represent the output of the system to another Gaussian input or to a deterministic input?

(P3) Assuming that the system θ when acting on deterministic inputs in, say, $L_2[a,b]$ has a Volterra kernel expansion

$$y(t) = k_0(t) + \sum_{p=1}^{\infty} L\int\cdots\int k_p(t;t_1,\ldots,t_p)x(t_1)\ldots x(t_p)dt_1\ldots dt_p \tag{3}$$

where L denotes Lebesgue integral, $y = \theta(x)$ and $k_p \epsilon L_2([a,b]^p)$, what is the relationship between the sets of kernels $\{f_p\}$, and $\{k_p\}$?

In the following we consider problems (P1) and (P3), which are straightforward, and we begin an exploration of the seemingly harder problem (P2). The analysis is based on the structure developed in [4].

We begin with problem (P1). Let $\{\phi_n\}$ be any complete orthonormal set in $\lambda_2(R)$. Then we have for every $t \epsilon T$ (omitting the arguments $t_1,\ldots,t_p$)

$$f_p(t) = \sum a_{n_1\ldots n_k}^{p_1\ldots p_k}(t)\, \phi_{n_1}^{\tilde{\otimes}p_1} \tilde{\otimes}\ldots\tilde{\otimes}\, \phi_{n_k}^{\tilde{\otimes}p_k} \tag{4}$$

where the sum is taken over all $k=1,\ldots,p$, $p_1+\ldots+p_k=p$, and $n_1,\ldots,n_k = 1,2,\ldots$ and converges in $\lambda_2(\otimes^p R)$, and the coefficients are given by

$$\begin{aligned} p_1!\ldots p_k!\,a_{n_1\ldots n_k}^{p_1\ldots p_k}(t) &= p!\,\langle f_p(t), \phi_{n_1}^{\tilde{\otimes}p_1} \tilde{\otimes}\ldots\tilde{\otimes}\, \phi_{n_k}^{\tilde{\otimes}p_k}\rangle_{\lambda_2(\tilde{\otimes}^p R)} \\ &= E\{J_p(f_p(t))J_p(\phi_{n_1}^{\tilde{\otimes}p_1} \tilde{\otimes}\ldots\tilde{\otimes}\, \phi_{n_k}^{\tilde{\otimes}p_k})\} \\ &= E\{Y_t H_{p_1}(\int\phi_{n_1}(s)X_s ds)\ldots H_{p_n}(\int\phi_{n_k}(s)X_s ds)\} \ . \end{aligned} \tag{5}$$

Hence from the input and output processes X and Y we can find the coefficients a from (5) and thus the kernels $f_p(t)$ from (4). Notice that the functions $\phi_n(t)$ can be chosen by orthonormalizing (in $\lambda_2(R)$ of course) any set of functions complete in $L_2(T)$; or else by putting $\phi_n = \lambda_n^{-\frac{1}{2}} e_n$ where $\{\lambda_n\}$ and $\{e_n\}$ are the eigenvalues and eigenfunctions of $R(t,s)$ [1]. This method of determining f_p has of course all the usual disadvantages of a series expansion. When $T = (-\infty,+\infty)$, X is stationary and $f_p(t;t_1,\ldots,t_p) = f_p(t_1-t, \ldots,t_p-t)$, $p=1,2,\ldots$, then a somewhat simpler method for approximating f_p can be found but we will not go into this here.

Problem (P2) is the most interesting as well as the most difficult one. Of course, if one is willing to put strong assumptions on the system, like those in problem (P3), then, as we shall see, the situation is fairly straightforward. Thus the main point of problem (P2) is to investigate conditions on the system, much weaker than those made in problem (P3), under which appropriate positive answers to problem (P2) can be given. Here we consider only the relationship of the representation of the system for the Gaussian input X to its representation for a deterministic input which is a sample function of X.

From (1) and (4) we have

$$\begin{aligned} Y_t - E(Y_t) &= \sum_{p=1}^{\infty} J_p(f_p) \\ &= \sum_{p=1}^{\infty} \sum_{\substack{k=1,\ldots,p \\ p_1+\ldots+p_k=p \\ n_1,\ldots,n_k=1,2,\ldots}} a_{n_1\ldots n_k}^{p_1\ldots p_k}(t) J_p\left(\phi_{n_1}^{\tilde{\otimes} p_1} \tilde{\otimes} \ldots \tilde{\otimes} \phi_{n_k}^{\tilde{\otimes} p_k}\right) . \end{aligned}$$

Then writing $H_p(\xi) = \sum_{j=0}^{p} c_j^p \xi^p$, for a standard normal random variable ξ, we have

$$\begin{aligned} J_p(\phi_{n_1}^{\tilde{\otimes} p_1} \tilde{\otimes} \ldots \tilde{\otimes} \phi_{n_k}^{\tilde{\otimes} p_k}) &= H_{p_1}(\xi_{n_1}) \ldots H_{p_k}(\xi_{n_k}) \\ &= \sum_{\substack{s_i=0,\ldots,p_i \\ i=1,\ldots,k}} c_{j_1}^{p_1} \ldots c_{j_k}^{p_k} \xi_{n_1}^{j_1} \ldots \xi_{n_k}^{j_k} \\ &= \sum_{\substack{j_i=0,\ldots,p_i \\ i=1,\ldots,k}} c_{j_1}^{p_1} \ldots c_{j_k}^{p_k} L\int\cdots\int(\phi_{n_1}^{\tilde{\otimes} j_1} \tilde{\otimes} \ldots \tilde{\otimes} \phi_{n_k}^{\tilde{\otimes} j_k})(t_1,\ldots,t_j) X_{t_1} \ldots X_{t_j} \\ &\qquad dt_1 \ldots dt_j , \end{aligned}$$

where $j=j_1+\ldots+j_k$, and if we define

$$h_q^N(t;t_1,\ldots,t_q) = \sum_{p=q}^{q+N} \sum_{\substack{k=1,\ldots,p \\ n_1,\ldots,n_k=1,\ldots,N}} \sum_{\substack{p_1+\ldots+p_k=p \\ j_1+\ldots+j_k=q}} c_{j_1}^{p_1} \ldots c_{j_k}^{p_k} a_{n_1\ldots n_k}^{p_1\ldots p_k}(t) \left(\phi_{n_1}^{\tilde{\otimes} j_1} \tilde{\otimes} \ldots \tilde{\otimes} \phi_{n_k}^{\tilde{\otimes} j_k}\right)(t_1,\ldots,t_q)$$

we obtain by a simple rearrangement of terms

$$Y_t - E(Y_t) = \sum_{q=0}^{\infty} \lim_{N\to\infty} L\int\cdots\int h_q^N(t;t_1,\ldots,t_q)X_{t_1}\ldots X_{t_q}dt_1\ldots dt_q \tag{6}$$

Since the convergence in (6) is in mean square, along some subsequence we will have convergence with probability one. Thus we can write

$$Y_t - E(Y_t) = \lim_{n\to\infty} \sum_{q=0}^{N_n} L\int\cdots\int h_q^{N_n}(t;t_1,\ldots,t_q)X_{t_1}\ldots X_{t_q}dt_1\ldots dt_q \quad \text{a.s.} \tag{7}$$

Thus for almost every sample function of the process X, the output of the system can be represented by (7). Notice that the kernels h_q^N in (7) can be found from the kernels f_p and that $h_q^N(t;t_1,\ldots,t_p)$ are continuous functions in $t_1,\ldots,t_p$ if we choose (as we can) the functions $\phi_n(t)$ to be continuous. Hence knowing the representation of the system output to a Gaussian input we can find the representation of the system output for a certain class of deterministic inputs, namely almost all sample functions of the Gaussian process. This deterministic input-output representation, given by (7), depends of course on the covariance R of the Gaussian process in the following two ways:

(i) R determines the kernels h_q^N of the representation (7), and

(ii) R determines, up to a zero probability set, the deterministic functions for which representation (7) is valid.

Representation (7) has the following two remarkable features. First, even though it is valid for a small class of deterministic inputs and its kernels depend on that class, it was obtained under extremely weak assumptions on the system, namely $E(Y_t^2) < \infty$. Second it is identical in form to the representation

$$y(t) = \lim_{n\to\infty}\{k_0^{N_n}(t) + \sum_{q=1}^{N_n} L\int\cdots\int k_q^{N_n}(t;t_1,\ldots,t_q)x(t_1)\ldots x(t_q)dt_1\ldots dt_q\} \tag{8}$$

obtained by Fréchet (1910) for all x(t) in C[a,b] or in $L_2[a,b]$ under the assumption that the system is continuous, in the sense that for each fixed t, y(t) is a continuous functional on C[a,b] or $L_2[a,b]$; in (8) the kernels k_q^N depend only on the system and not on the particular input x(t) in C[a,b] or in $L_2[a,b]$. It is thus remarkable that a representation like (8), valid not for all but only for some functions in $L_2[a,b]$ (or in C[a,b] if X has continuous sample functions), was derived with no continuity assumptions on the system.

If, furthermore, it turns out that for each q=1,2,..., the kernels $h_q^N(t)$ converge in $L_2([a,b]^q)$ to $h_q(t)$ (which constitutes an additional restriction on the system) then (6) and thus (7) may be simplified to

$$Y_t - E(Y_t) = \lim_{n\to\infty} \sum_{q=0}^{N_n} L\int\cdots\int h_q(t;t_1,\ldots,t_q)X_{t_1}\ldots X_{t_q}dt_1\ldots dt_q \quad \text{a.s.} \tag{9}$$

Finally, we should remark that if the system acting on the Gaussian input X is of finite order P, in the sense that in (1) $f_p = 0$ for $p > P$,

$$Y_t - E(Y_t) = \sum_{p=1}^{P} \int\cdots\int f_p(t;t_1,\ldots,t_p) X_{t_1}\ldots X_{t_p}\, dt_1\ldots dt_p \,, \tag{10}$$

then it has the same order P when acting on the sample functions of X, in the sense that (7) is written as

$$Y_t - E(Y_t) = \lim_{n\to\infty} \sum_{p=0}^{P} \int\cdots\int h_p^{N_n}(t;t_1,\ldots,t_p) X_{t_1}\ldots X_{t_p}\, dt_1\ldots dt_p \quad \text{a.s.} \tag{11}$$

and if, furthermore, each $h_p^N(t)$ converges in $L_2([a,b]^p)$ to $h_p(t)$ we have

$$Y_t - E(Y_t) = \sum_{p=0}^{P} \int\cdots\int h_p(t;t_1,\ldots,t_p) X_{t_1}\ldots X_{t_p}\, dt_1\ldots dt_p \quad \text{a.s.} \tag{12}$$

It may be checked (even though the calculations are somewhat messy) that in the latter case we always have

$$h_P = f_P \quad \text{and} \quad h_{P-1} = f_{P-1}$$

the assumption on the convergence of the $h_p^N(t)$'s implying that the kernels $f_p(t)$ belong to $L_2([a,b]^p)$.

We finally consider problem (P3) which consists in finding the kernels $\{k_p\}$ when the kernels $\{f_p\}$ are known, and vice versa. From (3) we have

$$Y_t = k_0(t) + \sum_{p=1}^{\infty} L\int\cdots\int k_p(t;t_1,\ldots,t_p) X_{t_1}\ldots X_{t_p}\, dt_1\ldots dt_p \quad \text{a.s.} \tag{13}$$

On the other hand representation (1) implies (7) which in view of (13) can be written as in (9). Now it follows from (9) and (13) that

$$k_0(t) = E(Y_t) + h_0(t)$$

and for p=1,2,...

$$L\int\cdots\int k_p(t;t_1,\ldots,t_p) X_{t_1}\ldots X_{t_p}\, dt_1\ldots dt_p$$
$$= L\int\cdots\int h_p(t;t_1,\ldots,t_p) X_{t_1}\ldots X_{t_p}\, dt_1\ldots dt_p \quad \text{a.s.} \tag{14}$$

If M_X^p is the subspace of $L_2([a,b]^p)$ generated by the symmetric functions $\{X_{t_1}(\omega)\ldots X_{t_p}(\omega),\ \omega\in\Omega-\Omega_0\}$ where Ω_0 is the exceptional set in (14), then (14) is equivalent to

$$h_p(t) = \text{Proj.}_{M_X^p} k_p(t)\ .$$

Thus, in general, knowledge of the kernels $\{f_p\}$, and thus the kernels $\{h_p\}$, determines only the projections of the kernels k_p onto M_X^p. The kernels k_p will be determined from the kernels f_p, by $k_p = h_p$, only if the subspaces

M_X^p consist of all symmetric functions in $L_2([a,b]^p)$. An equivalent condition is that if

$$R_p(t_1,\dots,t_p;s_1,\dots,s_p) = E(X_{t_1}\dots X_{t_p} X_{s_1}\dots X_{s_p})$$

and R_p denotes also the integral type operator in $L_2([a,b]^p)$ with kernel R_p, then R_p should be strictly positive definite, or the null space of $R_p^{\frac{1}{2}}$ should be $\{0\}$.

Conversely, knowledge of the kernels k_p clearly determines the kernels f_p. This is quite obvious from (13). The precise relationship can be derived by using the property

$$\begin{aligned}\int\cdots\int f_p(t;t_1,\dots,t_p)X_{t_1}\dots X_{t_p}\,dt_1\dots dt_p &= \mathrm{Proj}_{\overline{\mathcal{Q}}_p} Y_t \\ &= \sum_{q=p}^{\infty} L\int\cdots\int k_p(t;t_1,\dots,t_q)\,\mathrm{Proj}_{\overline{\mathcal{Q}}_p}(X_{t_1}\dots X_{t_q})\,dt_1\dots dt_q\end{aligned} \qquad (15)$$

where $\overline{\mathcal{Q}}_p$ is the closure in $L_2(X)$ of $\mathcal{Q}_p$, the set of all polynomials in the elements of $H(X)$ with degree p which are orthogonal to all polynomials with degree $\leq p-1$. The projections of $(X_{t_1}\dots X_{t_q})$ onto $\overline{\mathcal{Q}}_p$ can be expressed by using the Grad-Barrett-Hermite polynomials (see Section II of Root (1965)) and then (15) will give each f_p in terms of $k_p, k_{p+2},\dots$. When only the first P terms in (13) are present, the same will be true for (1) and in this case f_P can be expressed in terms of k_P, f_{P-1} in terms of k_{P-1}, f_{P-2} in terms of k_{P-2} and k_P, f_{P-3} in terms of k_{P-3} and k_{P-1}, etc. Again we will have in fact $f_P = k_P$ and $f_{P-1} = k_{P-1}$ as shown in Root (1965), where the specific expressions for $P = 3$ are also given.

ACKNOWLEDGMENT

This research was supported by the Air Force Office of Scientific Research under Grant AFOSR-75-2796.

REFERENCES

1. Cambanis, S. (1973). Representation of stochastic processes of second order and linear operations. *J. Math. Anal. Appl.* 41, 603-620.

2. Cameron, R.H. and Martin, W.T. (1947). The orthogonal development of nonlinear functionals in series of Fourier-Hermite functionals. *Ann. Math.* 48, 385-392.

3. Fréchet, M. (1910). Sur les fontionnelles continues. *Ann. Éc. Norm.* 27, 193-216.

4. Huang, S.T. and Cambanis, S. (1976). Stochastic and multiple Wiener integrals for Gaussian processes. *Institute of Statistics Mimeo Series No. 1087*, University of North Carolina at Chapel Hill.

5. Itô, K. (1944). Stochastic integral. *Proc. Imp. Acad. Tokyo*, 20, 519-524.

6. Itô, K. (1951). Multiple Wiener integrals. *J. Math. Soc. Japan*, 13, 157-169.

7. Root, W.L. (1965). On system measurement and identification. *Proceedings, Symposium on Systems Theory*, Polytechnic Institute of Brooklyn, Polytechnic Press, New York, 1965.

8. Wiener, N. (1938). The homogeneous chaos. *Amer. J. Math.*, 60, 897-936.

9. Wiener, N. (1958). *Nonlinear Problems in Random Theory*. Wiley, New York.

A REPRESENTATION FOR SPHERICALLY INVARIANT RANDOM PROCESSES

G.L. WISE
Department of Electrical Engineering
University of Texas at Austin
Austin, Texas 78712

N.C. GALLAGHER, JR.
School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907

ABSTRACT

It is shown that a random process is spherically invariant if and only if it is equivalent to a zero mean Gaussian process multiplied by an independent random variable. Several properties of spherically invariant random processes follow in a simple and direct fashion from this representation.

INTRODUCTION

Spherically invariant random processes (SIRP's) were introduced by Vershik [1] when he was investigating a class of random processes which shared some properties characteristic of Gaussian processes. In particular, he found that, for second order SIRP's, all mean square estimation problems have linear solutions, and this class of processes is closed under linear operations. In an interesting paper, Blake and Thomas [2] explored some important properties of SIRP's. Then, in a recent paper [3] Yao presented some very significant results concerning SIRP's. In particular, he presented a representation theorem for the family of finite dimensional distributions of SIRP's. The references in [3] provide a summary of other work done in this area.

By the well-known theorem of Kolmogorov [4; 5, pp.32-37], the consistent and symmetric family of all finite dimensional distributions of a random process is necessary and sufficient to give a statistical characterization of the process (i.e. to uniquely define the probability distribution in the sample function space for all Borel sets of the sample function space). In practice, few processes are characterized in this way; the major difficulty being the mathematical intractability of general n dimensional distributions. However, a SIRP can be conveniently characterized in terms of its family of finite dimensional distributions (see [2] and [3]). In this paper we will develop a representation theorem for SIRP's by using the family of finite dimensional distributions of a SIRP. It will be shown that, loosely speaking, at the heart of any SIRP is a Gaussian process. This will enable us to extend many results known for Gaussian processes to SIRP's.

DEVELOPMENT

We will say that a random process is singular if one random variable in the process can be written as a finite linear combination of the other random variables in the process. It will be seen later in this paper that (except for the identically zero process) the case considered by Yao is the nonsingular case. Thus the class considered by Yao does not include all Gaussian processes. Vershik, on the other hand, as well as Blake and Thomas, considered only second order SIRP's. We will impose neither the second order property nor the nonsingularity property.

For our purposes, a centered SIRP will be defined as a random process whose n-th order characteristic function is a function of an n-th order non negative definite quadratic form. If the mean exists, then the centering is easily seen to be equivalent to having a zero mean (i.e. use differentiation of the characteristic function). In the sequel it will be assumed that all SIRP's are centered. In Yao's work, the SIRP's were taken to be those whose n-th order characteristic function was a function of an n-th order positive definite quadratic form.

Let $\mathcal{C}$ denote the collection of all classes of consistent finite order characteristic functions of SIRP's whose n-th order characteristic function is a function of an n-th order positive definite quadratic form. Obviously, not just any function of the quadratic form will work. Let u_n represent an n dimensional row vector. Let R_n denote an nxn positive definite matrix. Let a prime denote the transpose. The following theorem is due to Yao [3].

Theorem 1: A necessary and sufficient condition for the class of characteristic functions

$$\{\varphi_n(u_n) = \psi_n(u_n R_n u_n'),\ n \geq 1\}$$

to be in $\mathcal{C}$ is that

$$\psi_n(r) = \psi(r) = \int_0^\infty \exp(-\frac{1}{2} r v^2)\, dF(v),$$

where $F(\cdot)$ is a probability distribution function supported on the non negative half line.

This theorem can be extended to include the singular case in the following straightforward manner. Consider a singular SIRP. Select N random variables from this process. Denote them by the row vector X_N. Assume that this set of random variables is singular. Let X_r denote the row vector of dimension r such that each random variable in X_r is also in X_N, X_r is nonsingular, and $X_N = X_r A$, where A is an rxN transformation matrix. Then from Theorem 1, we know that the characteristic function of X_r, denoted by

$$C_r(u_r) = E\{\exp(i\, u_r X_r')\}$$

is given by

$$C_r(u_r) = \int_0^\infty \exp(-\frac{v^2}{2} u_r R_r u_r')\, dF(v)$$

where R_r is some positive definite r×r matrix and $F(\cdot)$ is some probability distribution function supported on the non negative half line. The characteristic function of X_N is given by

$$\begin{aligned} C_N(u_N) &= E\{\exp(i u_N X_N')\} \\ &= E\{\exp(i u_N A' X_r')\} \\ &= C_r(u_N A') \\ &= \int_0^\infty \exp(-\frac{v^2}{2} u_N A' R_r A u_N')\, dF(v) . \end{aligned}$$

Thus we see that the characteristic function of X_N is a function of a non negative definite quadratic form.

Conversely, consider the class of characteristic functions

$$\left\{\varphi_n(u_n) = \int_0^\infty \exp(-\frac{v^2}{2} u_n \bar{R}_n u_n')\, dF(v) , \; n \geqslant 1 \right\}$$

where $\bar{R}_n$ is an n×n non negative definite matrix. It is readily seen that this is the class of characteristic functions of a random process which can be represented as a Gaussian random process, with zero mean and covariance matrix $\bar{R}_n$, multiplied by a random variable whose probability distribution function is given by $F(\cdot)$. Thus it follows from Kolmogorov's theorem that this class of characteristic functions is consistent. Also, from the definition of SIRP's, it follows that the corresponding random process is spherically invariant.

Let $\bar{\mathcal{C}}$ denote the collection of all classes of consistent finite order characteristic functions of SIRP's. Then we have the following extension of Theorem 1.

Theorem 2: Theorem 1 is true if $\mathcal{C}$ is replaced by $\bar{\mathcal{C}}$ and R_n is replaced by $\bar{R}_n$.

It follows from Theorem 2 and Kolmogorov's theorem that a random process is a SIRP if and only if it is of the form AY(t), where A is a random variable whose distribution function is supported on the non negative half line and

Y(t) is a zero mean Gaussian process independent of A. Consider, for the moment, the case where the distribution function of A is not necessarily supported on the non negative half line. The n-th order characteristic function of any n random variables in the process is of the form

$$C_n(u_n) = \int_{-\infty}^{\infty} \exp(-\frac{v^2}{2} u_n \bar{R}_n u_n') \, dF(v) ,$$

where $\bar{R}_n$ is the covariance matrix of the n mutually Gaussian random variables $Y(t_1), \ldots, Y(t_n)$. It is readily seen that this integral can be expressed as

$$C_n(u_n) = \int_{0}^{\infty} \exp(-\frac{v^2}{2} u_n \bar{R}_n u_n') \, d\tilde{F}(v) ,$$

where $\tilde{F}(\cdot)$ is a distribution function supported on the non negative half line. That is, if F is the distribution function of A, then $\tilde{F}$ is the distribution function of $|A|$. In other words, the random processes AY(t) and $|A|Y(t)$ have the same family of finite dimensional distributions. We summarize these results in the following representation theorem for SIRP's.

Theorem 3: A random process is a spherically invariant random process if and only if it is equivalent to a random process of the form AY(t), where A is an arbitrary random variable and Y(t) is a zero mean Gaussian process independent of A.

Theorem 3 characterizes SIRP's as zero mean Gaussian processes with random amplitudes. It is readily seen that a SIRP is Gaussian if and only if $F(\cdot)$ is atomic with only one atom (i.e. A is a constant).

In the sequel we will exclude the identically zero process from consideration. Consider the n-th order characteristic function associated with a SIRP. It follows in a straightforward fashion that the non negative definite matrix R_n in the quadratic form is the covariance matrix associated with the underlying Gaussian random variables, given in Theorem 3. If this matrix is positive definite, then it follows that the corresponding n Gaussian random variables are nonsingular. From this it follows that the n random variables from the SIRP are also nonsingular. In the case considered by Yao [3], R_n was positive definite for all n. Thus he considered nonsingular SIRP's. Conversely, if n random variables from a SIRP are singular, then the corresponding Gaussian random variables are singular, and thus the corresponding covariance matrix is not positive definite.

It is well known [1] that if a SIRP is ergodic, then it is a Gaussian process. Theorem 3 provides a simple illustration of this fact; that is, one sample function will not yield any information concerning the statistics of the random amplitude, and the amplitude is nonrandom if and only if the process is Gaussian.

Theorem 3 also provides a simple illustration of the fact [1] that a (deterministic) linear transformation of a SIRP results in a SIRP.

Theorem 3 furnishes a convenient method for studying the sample function properties of SIRP's. Sample function properties of Gaussian processes have been widely studied (see, for example, [5, Chapters 9-13]). We will consider a few special cases of interest, and we will state the results in terms of SIRP's. In the sequel, the zero mean SIRP $X(t)$ will be assumed to be separable, stationary, second order, and mean square continuous. Also, we will assume that $P\{X(t) = 0\} = 0$. The next theorem follows from the work of Dobrushin [6].

Theorem 4: Either

1. the sample functions of $X(t)$ are continuous with probability one

   or

2. with probability one, the sample functions of $X(t)$ have discontinuities of the second kind at every point.

The following corollary is a result of the further work of Belyaev [7].

Corollary 1: If the sample functions of $X(t)$ are not almost surely continuous, then they are almost surely unbounded in any interval.

The next theorem follows from the work of Hunt [8].

Theorem 5: Let $S(\omega)$ denote the spectral distribution function of $X(t)$. If

$$\int_0^\infty \left[\log(1+\omega)\right]^a dS(\omega) < \infty ,$$

for some $a > 1$, then $X(t)$ has continuous sample functions with probability one. Also, if

$$\int_0^\infty \omega^2 \left[\log(1+\omega)\right]^a dS(\omega) < \infty ,$$

for some $a > 1$, then $X(t)$ has continuous sample function derivatives with probability one.

The expected rate of zero crossings of a Gaussian random process has been studied by many investigators. The following theorem is based on the work of Ylvisaker [9].

Theorem 6: Let $R(\tau)$ denote the autocorrelation function of $X(t)$. The expected number of zero crossings of $X(t)$ in an interval of length T is given by

$$\frac{T}{\pi}\left[\frac{R''(0)}{R(0)}\right]^{1/2}$$

if $R(\tau)$ has a finite second derivative at the origin. If $R(\tau)$ does not have a finite second derivative at the origin, then the expected number of zero crossings in any interval is infinite.

Now we will consider some common zero memory nonlinearities and we will study their effects upon the second moment properties of $X(t)$. Represent $X(t)$ as $|A|Y(t)$, where $Y(t)$ is a Gaussian process independent of the random variable A. Also, take $Y(t)$ to have unit variance. Let $R(\tau)$ denote the autocorrelation function of $X(t)$, and let $\rho(\tau)$ denote the normalized autocorrelation function,

$$\rho(\tau) = \frac{R(\tau)}{R(0)} \quad .$$

Notice that $\rho(\tau)$ is also the autocorrelation function of $Y(t)$.

Let g_1 denote a hard limiter; that is,

$$g_1(u) = \begin{cases} 1 & \text{if } u \geqslant 0 \\ -1 & \text{if } u < 0 \end{cases} \quad .$$

It is clear that

$$E\{g_1[X(t+\tau)]\, g_1[X(t)]\} = E\{g_1[Y(t+\tau)]\, g_1[Y(t)]\} \quad .$$

Then it follows from the well known result for Gaussian processes [10, pp.280-284] that

$$E\{g_1[X(t+\tau)]\, g_1[X(t)]\} = \frac{2}{\pi} \text{ arc sin } \rho(\tau) \quad .$$

Let g_2 denote a half wave linear rectifier; that is

$$g_2(u) = \begin{cases} u & \text{if } u \geq 0 \\ 0 & \text{if } u < 0 \end{cases} \quad .$$

Then we have

$$E\{g_2[X(t+\tau)]\, g_2[X(t)]\} = E\Big\{E\{g_2[|A|Y(t+\tau)]\, g_2[|A|Y(t)] \mid A\}\Big\} .$$

It follows from [10, pp. 294-295] that

$$E\{g_2[|A|Y(t+\tau)]\, g_2[|A|Y(t)] \mid A\} = \frac{A^2}{4}\rho(\tau) + \frac{A^2}{2\pi}\Big[\rho(\tau) \arcsin \rho(\tau) + \sqrt{1-[\rho(\tau)]^2}\Big] .$$

Therefore, we have that

$$E\{g_2[X(t+\tau)]\, g_2[X(t)]\} = \frac{R(\tau)}{4} + \frac{R(0)}{2\pi}\Big[\rho(\tau) \arcsin \rho(\tau) + \sqrt{1-[\rho(\tau)]^2}\Big] .$$

Let g_3 denote the square law device; that is,

$$g_3(u) = u^2 \quad .$$

Assume that the fourth moment of X(t) exists. From the well known result for Gaussian processes [10, p. 264], we have

$$E\{g_3[|A|Y(t+\tau)]\, g_3[|A|Y(t)] \mid A\} = A^4 + 2A^4[\rho(\tau)]^2 \quad .$$

Then it follows that

$$E\{g_3[X(t+\tau)]\, g_3[X(t)]\} = [R(0)]^2 + 2[R(\tau)]^2 \quad .$$

Notice that for each of the above three zero memory nonlinearities, the transformation upon the second moment properties is exactly the same as for a Gaussian process. This result is not true in general. For example, consider Baum's limiter

$$g_4(u) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^u \exp(-\frac{v^2}{2})\, dv \ .$$

We have [10, pp.298-302]

$$E\{g_4[|A|Y(t+\tau)]\, g_4[|A|Y(t)]\ |A\} = \frac{1}{2\pi} \text{ arc sin} \left[\frac{|A|\rho(\tau)}{|A|+1}\right] .$$

Thus

$$E\{g_4[|A|Y(t+\tau)]\, g_4[Y(t)]\} = \frac{1}{2\pi} \int_0^\infty \text{arc sin} \left[\frac{v\rho(\tau)}{v+1}\right] dF(v) \ .$$

If

$$F(v) = \begin{cases} 0, & v < \sigma \\ 1, & v \geqslant \sigma \end{cases} ,$$

then X(t) is a Gaussian process with autocorrelation $\sigma^2\rho(\tau)$, and the above integral is equal to

$$\frac{1}{2\pi} \text{ arc sin} \left[\frac{\sigma\rho(\tau)}{\sigma+1}\right] .$$

If

$$F(v) = \begin{cases} 0, & v < \sigma/2 \\ 0.8, & \sigma/2 \leqslant v < 2\sigma \\ 1, & v \geqslant 2\sigma \end{cases} ,$$

then X(t) is a non-Gaussian SIRP with autocorrelation $\sigma^2\rho(\tau)$, and the above integral is equal to

$$\frac{2}{5\pi} \text{ arc sin} \left[\frac{\sigma\rho(\tau)}{\sigma+2}\right] + \frac{1}{10\pi} \text{ arc sin} \left[\frac{2\sigma\rho(\tau)}{2\sigma+1}\right] .$$

Thus the transformation upon the second moment properties induced by Baum's limiter when the input is a SIRP is not necessarily the same as when the input is a Gaussian process.

CONCLUSION

In this paper we have established a useful representation for SIRP's, which is given in Theorem 3. This theorem was then employed to establish several properties of SIRP's.

Consider a random process of the form AY(t), where Y(t) is a zero mean Gaussian process and A is an independent random variable. It is easy to show that such a random process is spherically invariant. However, the essential result of Theorem 3 is that any SIRP can always be represented by an equivalent random process having this form. This result illustrates the essential role played by Gaussian processes in the representation of SIRP's.

ACKNOWLEDGEMENTS

This research is supported by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant AFOSR-76-3062 and by the National Science Foundation under Grant ENG76-09443. The United States Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation hereon.

REFERENCES

1. A.M. Vershik, "Some characteristic properties of Gaussian stochastic processes," Theory of Prob. and Its Appl., vol. 9, pp.353-356, 1964.

2. I.F. Blake and J.B. Thomas, "On a class of processes arising in linear estimation theory," IEEE Trans. Info. Th., vol.IT-14, pp.12-16, Jan. 1968.

3. K. Yao, "A representation theorem and its applications to spherically-invariant random processes," IEEE Trans. Info. Th., vol.IT-19, pp.600-608, Sept. 1973.

4. A.N. Kolmogorov, Foundations of the Theory of Probability, Chelsea, New York, 1956, pp.27-33.

5. H. Cramér and M.R. Leadbetter, Stationary and Related Stochastic Processes, John Wiley, New York, 1967.

6. R.L. Dobrushin, "Properties of sample functions of a stationary Gaussian process," Theory of Prob. and Its Appl., vol.5, pp.120-121, 1960.

7. Yu.K. Belyaev, "Local properties of the sample functions of stationary Gaussian processes," Theory of Prob. and Its Appl., vol.5, pp.117-120, 1960.

8. G.A. Hunt, "Random Fourier transforms," Trans. Amer. Math. Soc., vol.71, pp.38-69, 1951.

9. N.D. Ylvisaker, "The expected number of zeros of a stationary Gaussian process," Ann. Math. Statist., vol.36, pp.1043-1046, 1965.

10. J.B. Thomas, An Introduction to Statistical Communication Theory, Wiley, New York, 1969.

ERROR-PROBABILITY BOUNDS FOR CONTINUOUS-TIME DECISION PROBLEMS[†]

J. L. HIBEY, D. L. SNYDER, AND J. H. VAN SCHUPPEN
Washington University, St. Louis, Missouri

SUMMARY

The computation of the probabilities of error arising in detection problems involving continuous-time stochastic processes is usually mathematically intractable. Thus, we address the problem of evaluating upper bounds for these error-probabilities. The approach is based on the application of martingale theory to detection and estimation problems. Using the state-variable approach of modelling non-linear systems, we have the following model. The signals are modulated by unobserved state processes that satisfy stochastic differential equations. The observation processes are either continuous or counting processes. In the first, the observation process contains the signal in additive noise. In the second, the signal is the rate process associated with a counting process, the details of which are given in [1].

Step 1. The decision strategy we employ is the well-known likelihood-ratio test, and the bounding technique is due to Chernoff. Starting with the generalized likelihood-ratio, we obtain an expression for the exponent of the Chernoff bound in terms of an expectation of a functional of the optimal signal estimates and, explicitly, the observation process.

Step 2. Next we perform a measure transformation to represent the expectation of Step 1 in terms of a simpler functional of estimates of the signals, simpler because it no longer depends explicitly on the observation process. We also obtain representations for the estimates of the signal and state processes and the observation process under this new measure. Evans [2] and Newman and Stuck [3] have also used measure transformations in evaluating error probability bounds.

Step 3. If, under the new measure, the estimates of the signals evolve as Markov processes, we are able to represent the expression of Step 2 as a solution to a backward Kolmogorov partial differential equation. When the observation process is discontinuous, the partial differential equation is a functional differential equation of the advanced type.

Step 4. We try to solve the partial differential equation.

With some modifications, the above procedure can also be used when the optimal signal estimates are replaced by suboptional estimates. These results are particularly important because the optimal filtering equations cannot always be solved exactly, numerical solutions are computationally infeasible, the optimal estimates are non-Markovian, and explicit systems for generating the exact estimates are known only in a few cases. We illustrate our procedure with examples that include the Gauss-Markov process and the doubly stochastic Poisson process.

References. [1] J. Hibey, "Error Prob. Bounds for Continuous-Time Decision Problems," D.Sc. Thesis, Wash. Univ., 1976. [2] J. Evans, IEEE Trans. on Info. Th., Sept. 1974. [3] C. Newman and B. Stuck, "Disc. Between Two Markov Processes," to be published.

[†]This work was partially supported by the National Science Foundation under Research Grants ENG 74-07800, ENG 76-11565, and ENG 76-10440 and by the National Institutes of Health under Research Grant Number RR00396 from the Division of Research Resources.

THE ZERO CROSSING INTERVAL STATISTICS OF THE SMOOTHED RANDOM TELEGRAPH SIGNAL

JOHN T. RICKARD
ORINCON Corporation
La Jolla, California

ABSTRACT

This paper presents explicit expressions for the zero crossing interval density, $P_0(\tau)$, of the smoothed random telegraph signal (SRTS).

I. INTRODUCTION

A basic problem in the theory of random processes is to determine the statistical features of the time intervals between successive zero crossings of a process. This area has been researched extensively since the pioneering work of Kac [1] and Rice [2], and many of the results are reviewed in the surveys by Levenbach [3] and Blake and Lindsey [4].

The probability density function $P_0(\tau)$ of the length of a zero crossing interval is perhaps the ultimate goal in a zero crossing analysis. However, discounting trivial cases (such as processes with Poisson zero crossing interval statistics), few attempts at obtaining closed-form expressions for $P_0(\tau)$ have met with any success. The Gaussian processes are a case in point. Wong [5], [6] considered the class of Gaussian processes with a correlation function $R(\tau)$ of the form

$$R(\tau) = 1 - \frac{1}{2}\tau^2 + c|\tau|^3 + o(|\tau|^3) ,$$

and was able to obtain and solve an integral equation for $P_0(\tau)$ when $c = 2\sqrt{3}/9$. This work has achieved some fame, since it represents the only case in which $P_0(\tau)$ has been determined explicitly for a Gaussian random process. Even greater difficulties are generally encountered with non-Gaussian processes, due in part to the mathematically intractable forms of their joint probability density functions. Nevertheless, the historical insolubility of the zero crossing problem has done little to diminish the interest in this subject.

In this paper, we present explicit results for the zero crossing interval statistics of the smoothed random telegraph signal (SRTS); i.e., the process occurring as the output of a low-pass RC filter which is excited by the random telegraph signal (Poisson square wave). Other properties of this process have been reported by Wonham and Fuller [7], Wonham [8], McFadden [9], Pawula [10], [11], and Rickard and Pawula [12]. The principal result of this paper is the derivation of $P_0(\tau)$ for the SRTS. We obtain an infinite series expression for $P_0(\tau)$ which, for integer values of a certain parameter of the process, reduces to a finite series. From these results, one may obtain the probability densities $P_n(\tau)$ of the sum of (n+1) successive zero crossing intervals, and the discrete probabilities $p(n, \tau)$ of having n zeroes in an arbitrarily chosen interval of length τ .

II. A RESULT ON THE ZERO CROSSINGS OF ARBITRARY PROCESSES

We begin by reviewing a result by McFadden [13], [14] on the zero crossing intervals of an arbitrary ergodic random process. Let $Z(t)$, $-\infty < t < \infty$, be such a process, and let $U(\tau)d\tau$ be the infinitesimal conditional probability that a zero crossing of $Z(t)$ occurs in the interval $(t_o + \tau,\ t_o + \tau + d\tau)$, given that a zero crossing occurred at time t_o (in the "horizontal window" sense of Kac and Slepian [15]). $U(\tau)$ can be written in terms of $P_n(\tau)$ as

$$U(\tau) = \sum_{n=0}^{\infty} P_n(\tau) \ , \tag{1}$$

since the latter crossing can be the first, second, or in general, the (n+1)st crossing after the one at t_o. Denoting the Laplace transforms of $U(\tau)$ and $P_n(\tau)$ by $u(s)$ and $p_n(s)$, respectively, we then have

$$u(s) = \sum_{n=0}^{\infty} p_n(s) \ . \tag{2}$$

If the zero crossing intervals of $Z(t)$ are statistically independent, then $P_n(\tau)$ is given by an n-fold convolution of $P_o(\tau)$ with itself, and thus it follows that $p_n(s) = [p_o(s)]^{n+1}$. Hence, for this case,

$$u(s) = \sum_{n=0}^{\infty} [p_o(s)]^{n+1} = \frac{p_o(s)}{1 - p_o(s)}$$

or

$$p_o(s) = \frac{u(s)}{1 + u(s)} \ . \tag{3}$$

Again, this result is valid only for processes having statistically independent zero crossing intervals.

III. THE SMOOTHED RANDOM TELEGRAPH SIGNAL

We now review some statistical features of the SRTS which are needed in later derivations. Let $X(t)$ be the random telegraph signal described by Kenrick [16] and Rice [2]. By definition, $X(t)$ takes only the values ± 1 with equal probability, and the number of transversals N of $X(t)$ in an interval $(0, T)$ is Poisson-distributed with parameter a. The SRTS, $Y(t)$, is generated as the output of a low-pass RC filter (with time constant $\beta^{-1} = RC$) excited by $X(t)$, and thus it satisfies the differential equation

$$\frac{d}{dt} Y(t) + \beta Y(t) = \beta X(t); \quad Y(t_o) = y_o \ , \tag{4}$$

at all points of continuity of X(t). Hence,

$$Y(t) = y_o \exp[-\beta(t-t_o)] + \int_{t_o}^{t} d\tau\, \beta \exp[-\beta(t-\tau)]\, X(\tau) \ . \tag{5}$$

Note that the SRTS is a two-parameter process, specified by the parameters a and β. Many of the statistical properties of the SRTS may be conveniently expressed in terms of the ratio $\alpha \triangleq a/\beta$. Two of these properties which we shall need are the mean zero crossing rate η_o, given by[1]

$$\eta_o = \frac{\beta}{B(\alpha, 1/2)} \ , \tag{6}$$

and the function $S(\tau)$, given by[2]

$$S(\tau) = \frac{\alpha\beta^2}{B(\alpha, 1/2)} \, {}_2F_1(1/2,\ 1-\alpha;\ 2;\ 1-\exp(-2\beta\tau)) \ , \tag{7}$$

for all $\alpha > 0$, $\beta > 0$, and $\tau \geq 0$. Note that for (positive) integer values of α, $1-\alpha$ is a nonpositive integer; hence, the right side of equation (7) is expressible as a finite summation having α terms. The above expression for η_o was first obtained by Pawula [11], and that for $S(\tau)$ was given by Rickard and Pawula [12].

As a final preliminary, it is easily verified, using (5) and the Poisson nature of X(t), that the zero crossing intervals of the SRTS are statistically independent. Hence, relation (3) of the previous section is valid for the SRTS.

---

[1] B(.,.) denotes the beta function.

[2] For any mean-square differentiable, stationary random process Z(t), $S(\tau)$ is defined by

$$S(\tau) \triangleq \int_{-\infty}^{\infty} d\xi \int_{-\infty}^{\infty} d\eta\, |\xi\eta|\, p(0, \xi;\ 0, \eta) \ ,$$

where $p(z_o, \dot{z}_o;\ z,\ \dot{z})$ is the joint density of the process and its derivative at time t and $t+\tau$, respectively [17]. The quantity $S(\tau)(d\tau)^2$ may be physically interpreted as the joint probability of zero crossings of Z(t) in both of the intervals $(t,\ t+d\tau)$ and $(t+\tau,\ t+\tau+d\tau)$. Here, ${}_2F_1(a, b;\ c;\ z)$ denotes the hypergeometric function.

IV. THE ZERO CROSSING INTERVAL DENSITY

Let us begin by deriving an explicit expression similar to equation (3) for the SRTS. We first note the close relationship between the function $U(\tau)$ defined in Section II and the function $S(\tau)$ in the previous section. Specifically, $U(\tau)d\tau$ is a conditional probability defined on the same two events as the joint probability $S(\tau)(d\tau)^2$, i.e., to $O(d\tau)$,

$$U(\tau)d\tau = \frac{S(\tau)(d\tau)^2}{\Pr\{\text{a zero crossing in } (t_o, t_o + d\tau)\}} \quad . \tag{8}$$

Now to $O(d\tau)$,

$$\Pr\{\text{a zero crossing in } (t_o, t_o + d\tau)\} = \eta_o d\tau = \frac{\beta d\tau}{B(\alpha, 1/2)} \quad , \tag{9}$$

where we have used (6). Substituting (7) and (9) into (8) gives

$$U(\tau) = \frac{\alpha\beta}{2} \, {}_2F_1(1/2,\ 1-\alpha;\ 2;\ 1-\exp(-2\beta\tau)) \quad , \tag{10}$$

for all $\alpha > 0$ and $\beta > 0$.

We now define a modified Laplace transform by the operator $L_\beta\{\cdot\}$, where

$$L_\beta\{f(t)\} \triangleq \int_0^\infty dt\ f(t) \exp(-2\beta\sigma\tau) \quad . \tag{11}$$

Obviously, this transform differs from the standard Laplace transform only by a change of scale, and thus equation (3) remains valid when the L_β-transforms are substituted for the Laplace transforms of the quantities in this equation. $L_\beta\{U(\tau)\}$ may be determined from (10) as follows:

$$\begin{aligned} L_\beta\{U(\tau)\} &= \frac{\alpha\beta}{2} \int_0^\infty dt \exp(-2\beta\sigma\tau)\ {}_2F_1(1/2,\ 1-\alpha;\ 2;\ 1-\exp(-2\beta\tau)) \\ &= -\frac{1}{2} + \frac{1}{2}(\sigma-1) \int_0^1 dz\ (1-z)^{\sigma-2}\ {}_2F_1(-1/2,\ -\alpha;\ 1;\ z) \quad , \end{aligned}$$

where the latter expression results from change of variables and integration by parts. The remaining integration is evaluated with the aid of [18] (p. 399, eq. 20.2 (4)), and we have

$$L_\beta\{U(\tau)\} = -\frac{1}{2} + \frac{1}{2}\,\frac{\Gamma(\sigma)\Gamma(\sigma+\alpha+1/2)}{\Gamma(\sigma+1/2)\Gamma(\sigma+\alpha)} \quad . \tag{12}$$

Substituting this result into (3) then gives

$$L_\beta\{P_o(\tau)\} = \frac{\Gamma(\sigma)\Gamma(\sigma+\alpha+1/2) - \Gamma(\sigma+1/2)\Gamma(\sigma+\alpha)}{\Gamma(\sigma)\Gamma(\sigma+\alpha+1/2) + \Gamma(\sigma+1/2)\Gamma(\sigma+\alpha)} \quad , \tag{13}$$

which is the desired expression.

The zero crossing interval density $P_o(\tau)$ may be found by using contour integration and the residue theorem to invert (13). We shall first demonstrate the solution for the general case where α takes on arbitrary (positive) values, and then show how this result simplifies when α assumes integer values.

The inverse L_β-transform of (13) is given by

$$P_o(\tau) = \frac{\beta}{i\pi}\int_{c-i\infty}^{c+i\infty} d\sigma \exp(2\beta\sigma\tau)\, \frac{\Gamma(\sigma)\Gamma(\sigma+\alpha+1/2) - \Gamma(\sigma+1/2)\Gamma(\sigma+\alpha)}{\Gamma(\sigma)\Gamma(\sigma+\alpha+1/2) + \Gamma(\sigma+1/2)\Gamma(\sigma+\alpha)} \qquad (14)$$

where the line $\sigma = c$ lies in the right half of the complex plane. By inspection, the singularities of the integrand in (14) correspond precisely to the zeroes of the denominator, i. e., the roots of the transcendental equation

$$\Gamma(\sigma)\Gamma(\sigma+\alpha+1/2) + \Gamma(\sigma+1/2)\Gamma(\sigma+\alpha) = 0 \ . \qquad (15)$$

It may be shown (see the Appendix for details) that the roots of this equation are distinct, and are isolated into disjoint intervals along the negative real axis. Hence, their computation (using Newton's method, for example) is a simple exercise, and they may be tabulated, as is done for the zeroes of Bessel functions and other transcendental functions. Denoting these roots by σ_j, $j = 1, 2, \ldots$, such that $\sigma_1 > \sigma_2 > \ldots$, the residues $r(\sigma_j)$ of the integrand in (14) may be calculated by employing l'Hospitals' rule to yield[1]

$$r(\sigma_j) = \frac{2\beta}{i\pi}\left[\psi(\sigma_j) - \psi(\sigma_j+\alpha) + \psi(\sigma_j+\alpha+1/2) - \psi(\sigma_j+1/2)\right]^{-1} \exp(2\beta\sigma_j\tau)$$

Hence, from (14) and the residue theorem, $P_o(\tau)$ is given by $2\pi i$ times the sum of the $r(\sigma_j)$; viz.,

$$P_o(\tau) = \sum_{j=1}^{\infty} q(\sigma_j)\exp(2\beta\sigma_j\tau) \ , \qquad (16)$$

where

$$q(\sigma_j) \triangleq 4\beta\left[\psi(\sigma_j) - \psi(\sigma_j+\alpha) + \psi(\sigma_j+\alpha+1/2) - \psi(\sigma_j+1/2)\right]^{-1} \ . \qquad (17)$$

This expression is valid for all parameter values of the SRTS.

We now consider in some detail the special case where the parameter α assumes integer values. To avoid confusion, we shall use the variable m to represent integer values of α in the discussion which follows.

---

[1] $\psi(z)$ denotes the logarithmic derivative of the gamma function ([19], p. 15); viz.,

$$\psi(z) \triangleq \frac{d}{dz}\ln\Gamma(z) = \frac{1}{\Gamma(z)}\frac{d}{dz}\Gamma(z) \ .$$

When $\alpha = m$, an integer, equation (15) reduces to the m^{th}-order polynomial equation[1]

$$(\sigma + 1/2)_m + \sigma_m = 0 \; . \tag{18}$$

This equation has precisely m roots, one each in the right half of the first m unit intervals along the negative real axis (see the Appendix). Given this property, the numerical computation of the roots is quite simple; and, in fact, by using the symmetry property given in the Appendix, we can obtain exact values for m = 1 through 9 by employing the well-known root formulas for polynomials up to the fourth degree. Table 1 lists the roots for m = 1 to 5.

The zero crossing interval density, equation (16), also simplifies considerably when α is an integer. Since ([19], p. 16)

$$\psi(z + n) = \psi(z) + \sum_{k=0}^{n-1} \frac{1}{z+k} \; ,$$

we have from equation (17) when $\alpha = m$, an integer,

$$q(\sigma_j) = 8\beta \left[\sum_{k=0}^{m-1} \frac{-1}{(\sigma_j + k)(\sigma_j + k + 1/2)} \right]^{-1}$$

Thus, the interval density for integer α is given by

$$P_o(\tau) = 8\beta \sum_{j=1}^{m} \left[\sum_{k=0}^{m-1} \frac{-1}{(\sigma_j + k)(\sigma_j + k + 1/2)} \right]^{-1} \exp(2\beta\sigma_j\tau) \; . \tag{19}$$

Using Table 1 and equation (19) for the cases $\alpha = 1$, 2, and 3, yields the following:

$$P_o(\tau) = \begin{cases} \frac{\beta}{2} \exp\left(-\frac{1}{2}\beta\tau\right) & \alpha = 1 \; , \\ \beta \exp\left(-\frac{3}{2}\beta\tau\right) \cosh\left(\frac{\sqrt{3}}{2}\beta\tau\right) \; , & \alpha = 2 \; , \\ \frac{3\beta}{26} \exp\left(-\frac{5}{2}\beta\tau\right)\left[5 + 8\cosh\left(\frac{\sqrt{13}}{2}\beta\tau\right)\right] \; , & \alpha = 3 \; . \end{cases}$$

These are exact expressions for the interval density with the given values of α.

---

[1] The notation $(a)_k$ denotes the Pochhammer symbol:

$$(a)_k = a(a+1) \ldots (a + k - 1) \; .$$

Figure 1 shows curves of $P_o(\tau)$ over a range of values for the parameter α. In this figure, the vertical and horizontal axes have been scaled by factors of $(2\beta)^{-1}$ and 2β, respectively.

As a final note, we observe that $P_n(\tau)$ $(n \geq 1)$ is given by an n-fold convolution of $P_o(\tau)$ with itself. Furthermore, by using the recursion relations of McFadden [14], we may obtain $p(n, \tau)$ $(n \geq 0)$. The details are given in [20].

ACKNOWLEDGEMENT

The author expresses his gratitude to R. Pawula and S. O. Rice for their encouragement and helpful comments during the course of this work. This research was accomplished under the auspices of the Society for Statistical Geometry.

APPENDIX

The roots of equation (15) satisfy certain properties which greatly ease their computational burden. The proofs of these properties are straightforward but somewhat tedious. Hence, they are not included here, but may be found in [20].

For non-integer values of α, the following properties hold:

Property A1. The roots σ_j, $j = 1, 2, \ldots$, of (15) are distinct, and lie along the negative real axis.

Property A2. Let the roots σ_j, $j = 1, 2, \ldots$, be ordered so that $\sigma_1 > \sigma_2 > \ldots$. Furthermore, let $\alpha = M + \delta$, where M is an integer and $0 < \delta < 1$. Then for $j = 1, 2, \ldots, M$, $\sigma_j \in (-j + 1/2, -j + 1)$. The remaining roots σ_j, $j = M+1, M+2, \ldots$, are isolated, one per interval, according to the following cases:

1. $0 < \delta \leq 1/2$. The roots lie in the intervals $(-\alpha-k, -M-k)$ and $(-\alpha-k-1/2, -M-k-1/2)$, $k = 0, 1, \ldots$.
2. $1/2 < \delta < 1$. The roots lie in the intervals $(-M-k-1, -\alpha-k)$ and $(-M-k-3/2, -\alpha-k-1/2)$, $k = 0, 1, \ldots$.

With the aid of these properties, the roots σ_j, $j = 1, 2, \ldots$, of (15) are easily computed using, for example, Newton's method.

Even greater simplifications result when $\alpha = m$, an integer. We now seek the roots of equation (18), which satisfy (see [20] for proofs):

Property I1. The roots σ_j, $j = 1, 2, \ldots, m$, of (18) are distinct, and lie along the negative real axis. Furthermore, if the σ_j are ordered so that $\sigma_1 > \sigma_2 > \ldots > \sigma_m$, then $\sigma_j \in (-j + 1/2, -j + 1)$.

Property I2. With the above ordering, if m is odd, then the middle root $\sigma_{\frac{m+1}{2}}$ is given by

$$\sigma_{\frac{m+1}{2}} = -\frac{2m-1}{4}$$

For any m, if σ_j, $j \leq m/2$, is a root of (18), then

$$\sigma_{m-j+1} = -\left(\sigma_j + \frac{2m-1}{2}\right), \; j \leq m/2 \; .$$

Hence, only half of the roots in the integer α case need be computed numerically.

REFERENCES

1. M. Kac, "On the average number of real roots of a random algebraic equation," Bull. Amer. Math. Soc., vol. 49, pp. 314-319, 1943.

2. S. O. Rice, "The mathematical analysis of random noise," Bell Syst. Tech. J., vol. 23, pp. 282-332, 1944. Also, vol. 24, pp. 46-156, 1945.

3. H. Levenbach, "The zero-crossing problem," Queen's Univ., Kingston, Ont., Res. Rept. 63-4, 1963.

4. I. F. Blake and W. C. Lindsey, "Level-crossing problems for random processes," IEEE Trans. Inform. Theory, vol. IT-19, pp. 295-315, 1973.

5. E. Wong, "Some results concerning the zero crossings of Gaussian noise," SIAM J. Appl. Math., vol. 14, pp. 1246-1254, 1966.

6. E. Wong, "The distribution of intervals between zeroes for a stationary Gaussian random process," SIAM J. Appl. Math., vol. 18, pp. 67-73, 1970.

7. W. M. Wonham and A. T. Fuller, "Probability densities of the smoothed random telegraph signal," J. Electr. Contr., vol. 4, pp. 567-576, 1958.

8. W. M. Wonham, "Transition probability densities of the smoothed random telegraph signal," J. Electr. Contr., vol. 6, pp. 376-384, 1959.

9. J. A. McFadden, "The probability density of the output of an RC filter when the input is a binary random process," IRE Trans. Inform. Theory, vol. IT-5, pp. 174-178, 1959.

10. R. F. Pawula, "The transition probability density function of the low-pass filtered random telegraph signal," Inter. J. Contr., vol. 12, pp. 25-32, 1970.

11. R. F. Pawula, "Statistical geometry of the smoothed random telegraph signal," Inter. J. Contr., vol. 16, pp. 629-640, 1972.

12. J. T. Rickard and R. F. Pawula, "The zero crossing variance of the smoothed random telegraph signal," Inter. J. Contr., vol. 21, pp. 743-752, 1975.

13. J. A. McFadden, "The axis-crossing intervals of random functions," IRE Trans. Inform. Theory, vol. IT-2, pp. 146-150, 1956.

14. J. A. McFadden, "The axis-crossing intervals of random functions-II," IRE Trans. Inform. Theory, vol. IT-4, pp. 14-24, 1958.

15. M. Kac and D. Slepian, "Large excursions of Gaussian processes," Ann. Math. Statist., vol. 30, pp. 1215-1228, 1959.

16. G. W. Kenrick, "The analysis of irregular motions with applications to the energy frequency spectrum of static and of telegraph signals," Phil. Mag., ser. 7, vol. 7, pp. 176-196, 1929.

17. H. Steinberg, P. M. Schultheiss, C. A. Wogrin, and F. Zweig, "Short-time frequency measurement of narrow-band random signals by means of a zero counting process," J. Appl. Phys., vol. 26, pp. 195-201, 1955.

18. A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, Tables of Integral Transforms, Vol. 2, New York: McGraw-Hill, 1954.

19. A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, Higher Transcendental Functions, Vol. 1, New York: McGraw-Hill, 1953.

20. J. T. Rickard, "The zero crossing properties of the smoothed random telegraph signal," Ph. D. dissertation, Univ. of Calif., San Diego, 1975.

Table 1. Roots of Equation (18) for m = 1 through 5

m	σ_1	σ_2	σ_3	σ_4	σ_5
1	$-\frac{1}{4}$				
2	$\frac{-3+\sqrt{3}}{4}$	$\frac{-3-\sqrt{3}}{4}$			
3	$\frac{-5+\sqrt{13}}{4}$	$-\frac{5}{4}$	$\frac{-5-\sqrt{13}}{4}$		
4	$\frac{-7+\sqrt{17+2\sqrt{46}}}{4}$	$\frac{-7+\sqrt{17-2\sqrt{46}}}{4}$	$\frac{-7-\sqrt{17-2\sqrt{46}}}{4}$	$\frac{-7-\sqrt{17+2\sqrt{46}}}{4}$	
5	$\frac{-9+\sqrt{35+2\sqrt{109}}}{4}$	$\frac{-9+\sqrt{35-2\sqrt{109}}}{4}$	$-\frac{9}{4}$	$\frac{-9-\sqrt{35-2\sqrt{109}}}{4}$	$\frac{-9-\sqrt{35+2\sqrt{109}}}{4}$

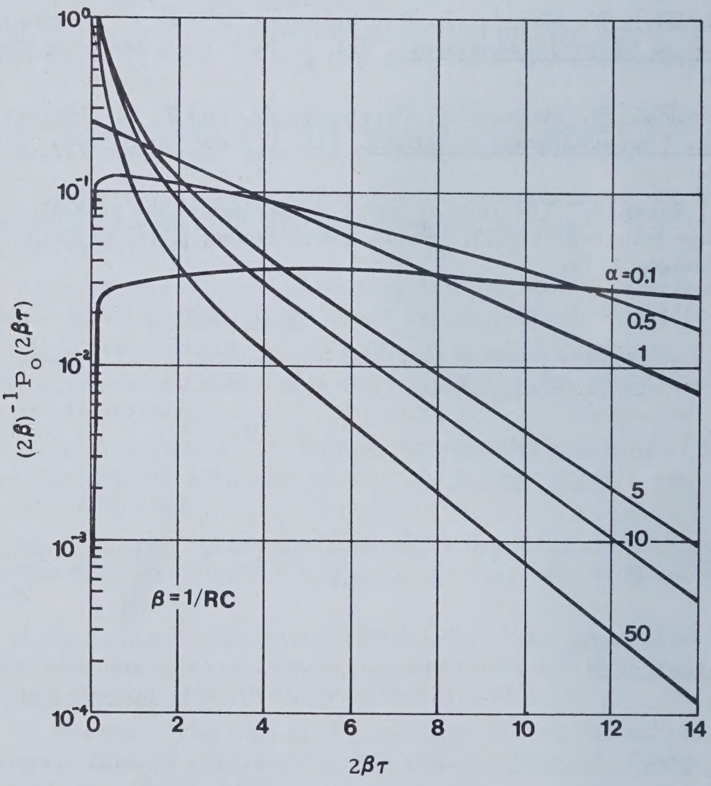


Figure 1. Interval Density $P_o(\tau)$

POWER MEAN FREQUENCY ESTIMATION FOR A RANDOM PROCESS GATED AT A SUB-NYQUIST RATE

JAYVADAN N. PATEL AND LESTER C. PEACH
Laboratory for Atmospheric Probing
University of Chicago and Illinois Institute of Technology

SUMMARY

The estimation of the power mean frequency of a narrowband, stationary, ergodic random process by processing a sample function of the random process is well explored subject.[1,2] Its application is found in radar meteorology where the power spectral density of the doppler shifted radar return is related to the velocity distribution of meteorological scatterers in a scattering volume.

This paper investigates the case when a gated sample function of the aforementioned random process is available for processing and the frequency of the gating function is lower than the Nyquist rate required for the narrowband process. This is encountered in the use of a pulsed radar for meteorological measurements when the spread of the doppler spectrum is larger than one half of the pulse repetition frequency (PRF) of the radar. The Sub-Nyquist gating rate does not interfere with the estimation of the power mean frequency of the random process. The estimator performance is identical to the estimators for CW signal when the integration time is infinite. A modification is required to apply the developed theory to the pulsed radar return signal due to meteorological scatterers. The smearing of returns from adjacent scattering volumes requires these modification. The modification involves the representation of an integral by a sum of samples of the integrand, where the sampling rate is Sub-Nyquist. The finite time integration does not bias the estimator. However, the variance of the estimator is increased due to the Sub-Nyquist sampling during the finite time.

The feasibility of the proposed estimator can be explained as follows. The Sub-Nyquist gating rate is not a hinderance as the gated signal is processed before the final Sub-Nyquist sampling and summation. The Nyquist rate restriction will occur if the gated signal is sampled at the gating frequency and the samples then processed. Thus processing prior to sampling allows the successful processing of signals gated at a Sub-Nyquist rate.

ACKNOWLEDGEMENT

This research was supported in part by National Science Foundation grant NSF-DES-75-09683.

REFERENCES

1. Denenberg J. N., 1971, The Estimation of Spectral Moments; Ph.D.Dissertation, Electrical Engineering Department, Illinois Institute of Technology, Chicago.

2. Sirmens, D. and Bill Bumgarner, 1975, Numerical Comparison of five Mean Frequency Estimators; Journal of Applied Meteorology, Sept. 1975, vol. 14, No. 6, pp. 991-1003.

SIMULATION OF ABNORMAL CELL BEHAVIOUR

E.J. DAVISON

Dept. of Electrical Engineering
University of Toronto
Toronto, Canada M5S 1A4

ABSTRACT

A mathematical model of a cell growing in a culture medium was obtained in [1]. Using this model, various behaviour patterns of the cell were obtained under different types of disturbances. In particular, conditions were found which caused the cell to behave in a rapidly growing 'tumor-like' mode. This paper describes some of the properties of the abnormal cell behaviour obtained in [1] in greater depth. In particular it is shown that the 'tumor -like' cell obtained is relatively insensitive to parameter perturbations unlike the normal cell. In addition, it is shown that severe ionizing radiation can sometimes(not always) cause the 'tumor-like' cell to revert back to its previous normal behaviour.

1. INTRODUCTION

In [1], a mathematical model of a cell growing in a culture medium was obtained, and conditions which caused the cell to behave in an abnormal way were obtained by systematically applying various disturbances to the cell. It was found that some disturbances cause little change in the cell's behaviour, some cause the cell to die, and other disturbances cause the cell to behave in a 'tumor-like' manner, i.e. the cell's division time is greatly speeded up, with the size of the cell being much larger than for a normal cell. It is the purpose of this paper to describe some of the properties of this 'tumor-like' cell in greater depth.

2. NORMAL CELL MODEL

This section deals with the qualitative behaviour assumed for the model of a cell growing in a culture medium.

2.1 Normal Cell Chemistry

The basic components (states) of the cell system are: genes , messengers, templates, enzymes, repressors and activators. The following biochemical equations (Table I), taken from Heinments [2], are assumed to describe the cell's behaviour. (See end of paper for definition of symbols used.) It is observed that all dynamic behaviour outside the nucleus of the cell has been ignored in this model and that the extra-cellular nutrient pool P_e is assumed to be a constant (but its value can be varied). It is in addition assumed that the temperature of the cell

---

This work has been supported by the National Research Council under Grant No. A4396.

TABLE I

Complete Set of Reactions Occurring in the Cell

1. $E_p + P_n \xrightarrow{k_1} [E_pP_n]$
2. $B' + E \xrightarrow{k_2} B$
3. $G_B + P_n \xrightarrow{k_3} G_B + B'$
4. $G_C + P_n \xrightarrow{k_4} G_P + M_p$
5. $G_p + P_n \xrightarrow{k_5} G_p + M_p$
6. $G_E + [E_pP_n] \xrightarrow{k_6} G_E + E_p + M$
7. $B + M \xrightarrow{k_7} N$
8. $B + M_p \xrightarrow{k_8} N_p$
9. $C + P_a \xrightarrow{k_9} [CP_a]$
10. $N + [CP_a] \xrightarrow{k_{10}} B + C + M + E$
11. $N_p + [CP_a] \xrightarrow{k_{11}} B + M_p + C + E_p$
12. $M \xrightarrow{k_{12}} P_n$
13. $M_p \xrightarrow{k_{13}} P_n$
14. $B \xrightarrow{k_{14}} P_i$
15. $C \xrightarrow{k_{15}} P_i$
16. $E \xrightarrow{k_{16}} P_i$
17. $E_p \xrightarrow{k_{17}} P_i$
18. $P_e + E \xrightarrow{k_{18}} P_i + E$
19. $P_i + E \xrightarrow{k_{19}} P_n + E$
20. $P_i + E \xrightarrow{k_{20}} P_a + E$
21. $E_p + C \underset{k_{31}}{\overset{k_{21}}{\rightleftharpoons}} [E_pC]$
22. $E + P_i \underset{k_{32}}{\overset{k_{22}}{\rightleftharpoons}} [EP_i]$
23. $E_p + B \underset{k_{33}}{\overset{k_{23}}{\rightleftharpoons}} [E_pB]$
24. $N \xrightarrow{k_{28}} P_i$
25. $N_p \xrightarrow{k_{29}} P_i$
26. $P_i \xrightarrow{k_{30}} X$

is held fixed. Figure 1 gives a block diagram for the interconnections of the various components of the cell.

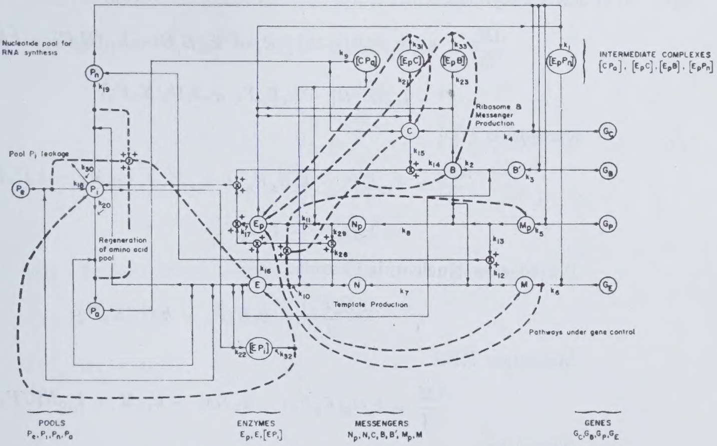

Nucleotide pool for RNA synthesis
Pool P_i leakage
Regeneration of amino acid pool
Ribosome & Messenger Production
Template Production
Pathways under gene control
INTERMEDIATE COMPLEXES
[CP_a], [E_pC], [E_pB], [E_pP_n]
P_n
P_i
P_e
P_a
E_p
E
[EP_i]
N_p
N
C
B
B'
M_p
M
[CP_a]
[E_pC]
[E_pB]
[E_pP_n]
G_C
G_B
G_P
G_E
POOLS
P_e, P_i, P_n, P_a
ENZYMES
E_p, E, [EP_i]
MESSENGERS
N_p, N, C, B, B', M_p, M
GENES
G_C, G_B, G_P, G_E

Figure 1. Flow chart showing dynamics of a cell nucleus (see Table 1)

2.2 Mathematical Model of Cell. The following equations describe the mathematical model of the cell assumed in the paper; they are directly obtained from the chemical reaction equations of Table I using standard mass balance techniques:

RNA Fraction of Ribosome

$$\frac{\mathrm{d}B'}{\mathrm{d}t} = k_3 G_B P_N - k_2 B'E \tag{1}$$

Ribosome

$$\begin{aligned}\frac{\mathrm{d}B}{\mathrm{d}t} &= k_2 B'E - k_7 BM - k_8 BM_p - k_{14} B + k_{10} N[CP_a] \\ &\quad + k_{11} N_p[CP_a] - k_{23} E_p B + k_{33}[E_p B]\end{aligned} \tag{2}$$

Transport RNA

$$\begin{aligned}\frac{\mathrm{d}C}{\mathrm{d}t} &= k_4 G_C P_n - k_9 CP_a + k_{10} N[CP_a] + k_{11} N_p[CP_a] - k_{15} C \\ &\quad - k_{21} E_p C + k_{31}[E_p C]\end{aligned} \tag{3}$$

General Intercellular Metabolic Pool

$$\begin{aligned}\frac{\mathrm{d}P_i}{\mathrm{d}t} &= k_{18} EP_e - k_{19} EP_i - k_{20} EP_i + k_{32}[EP_i] + k_{14} B - k_{22} EP_i \\ &\quad + k_{15} C + k_{16} E + k_{17} E_p + k_{28} N + k_{29} N_p - k_{30} P_i\end{aligned} \tag{4}$$

Polymerase–Transport RNA Complex

$$\frac{\mathrm{d}[E_p C]}{\mathrm{d}t} = k_{21} E_p C - k_{31}[E_p C] \tag{5}$$

Polymerase–RNA Complex

$$\frac{\mathrm{d}[E_p B]}{\mathrm{d}t} = k_{23} E_p B - k_{33}[E_p B] \tag{6}$$

RNA Polymerase

$$\begin{aligned}\frac{\mathrm{d}E_p}{\mathrm{d}t} &= k_{11} N_p[CP_a] - k_{17} E_p - k_{21} E_p C + k_{31}[E_p C] - k_{23} E_p B \\ &\quad + k_{33}[E_p B] - k_1 E_p P_n + k_6 G_E[E_p P_n]\end{aligned} \tag{7}$$

Nucleotide Pool

$$\begin{aligned}\frac{\mathrm{d}P_n}{\mathrm{d}t} &= k_{19} EP_i - k_1 E_p P_n + k_{12} M + k_{13} M_p - k_4 G_C P_n \\ &\quad - k_5 G_p P_n - k_3 G_B P_n\end{aligned} \tag{8}$$

Polymerase–Nucleotide Complex

$$\frac{\mathrm{d}[E_p P_n]}{\mathrm{d}t} = k_1 E_p P_n - k_6 G_E[E_p P_n] \tag{9}$$

Messenger RNA

$$\frac{\mathrm{d}M}{\mathrm{d}t} = k_6 G_E[E_p P_n] - k_7 BM - k_{12} M + k_{10} N[CP_a] \tag{10}$$

Template

$$\frac{\mathrm{d}N}{\mathrm{d}t} = k_7 BM - k_{10} N[CP_a] - k_{28} N \tag{11}$$

Amino Acid Pool

$$\frac{dP_a}{dt} = k_{20}P_tE - k_9CP_a \tag{12}$$

Transport RNA–Amino Acid Complex

$$\frac{d[CP_a]}{dt} = k_9CP_a - k_{10}N[CP_a] - k_{11}N_p[CP_a] \tag{13}$$

Total Protein

$$\frac{dE}{dt} = k_{10}N[CP_a] - k_{16}E - k_{22}EP_i + k_{32}[EP_i] - k_2EB' \tag{14}$$

Protein–Intercellular Metabolic Complex

$$\frac{d[EP_i]}{dt} = k_{22}EP_i - k_{32}[EP_i] \tag{15}$$

RNA Polymerase Cycle

Messenger RNA

$$\frac{dM_p}{dt} = k_5G_pP_n - k_8BM_p + k_{11}N_p[CP_a] - k_{13}M_p \tag{16}$$

Template

$$\frac{dN_p}{dt} = k_8BM_p - k_{11}N_p[CP_a] - k_{29}N_p \tag{17}$$

State equations of the cell. For the sake of conciseness, the above equations are described by:

$$\dot{x} = f(x, u, k), \quad x(0) = x^\circ, \quad 0 \leq t \leq t^*, \tag{18}$$

a set of 17 simultaneous differential equations of the quadratic type, where

$x \in R^{17}$ is the state of the cell given by:

$$\begin{array}{lll} x_1 = E_p & x_7 = M & x_{13} = E \\ x_2 = P_n & x_8 = M_p & x_{14} = P_i \\ x_3 = [E_pP_n] & x_9 = N & x_{15} = [E_pC] \\ x_4 = B' & x_{10} = N_p & x_{16} = [E_pB] \\ x_5 = B & x_{11} = P_a & x_{17} = [EP_i] \\ x_6 = C & x_{12} = [CP_a] & \end{array} \tag{19}$$

where $u \in R^5$ are the external inputs to the cell given by:

$u_1 = P_e \quad u_3 = G_c \quad u_5 = G_p$

$u_2 = G_E \quad u_4 = G_B$

where $k \in R^{29}$ is a vector of rate constants for the cell and $x(0) \in R^{17}$ are the initial conditions of the cell, t^* is the division time of the cell, at which time cell division takes place. The above equations (19) describe the behaviour of a single cell from the time of birth to the time at which cell division takes place (t^* seconds later).

2.3 Mathematical Model of a Growing Cell. To describe the behaviour of a growing cell, it is necessary to introduce some type of mitosis hypothesis (since the biological details as to when a cell divides are not well understood). Consider the system (18); the following division hypothesis is made:

Division hypothesis. Let $t^* \in [0, \infty)$ be the time which globally minimizes the criterion function $J(t) \triangleq \| x(t) - 2x(0) \|$. Then if $J(t^*) < \| x(0) \|$, cell division takes place at $t = t^*$; if $J(t^*) \geq \| x(0) \|$, cell division does not take place. If cell division takes place, the initial condition of the new generated cell is equal to $\frac{1}{2}x(t^*)$. Figure 2 illustrates this division hypothesis.

The motivation for such a hypothesis is that a cell in steady-state behaviour, i.e. with identical dynamic behaviour existing from one generation of cell growth to another, will have the property that $J(t^*) = 0 < \| x(0) \|$, and $x(0) = \frac{1}{2}x(t^*)$.

Let x_i, t_i^*, $i = 1,2,3,\ldots$ denote the state and division time respectively of

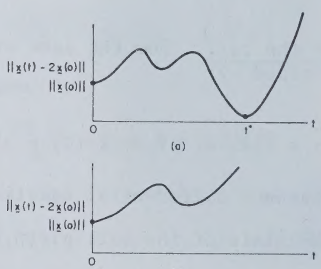


Figure 2. Criterion for deciding when mitosis takes place. (a) Cell divides at time t^* in this case. (b) Cell does not divide in this case

the parent cell, 1st daughter cell, 2nd daughter cell etc. Then the behaviour of a growing cell is thus described by the following sequence of differntial equations:

$$\dot{x}_1 = f(x_1, u, k), \quad x_1(0) = x^\circ, \quad 0 \leq t \leq t^*_1. \qquad (21)$$

$$\dot{x}_{i+1} = f(x_{i+1}, u, k), \quad x_{i+1}(0) = \tfrac{1}{2}x_i(t_i^*), \quad 0 \leq t \leq t^*_{i+1} \quad i = 1,2,3, \ldots$$

where t_i^*, i = 1,2,3,... is found so that $\min_{t_i \in [0,\infty)} || x_i(t_i) - 2x_i(0) ||$ occurs at $t_i = t_i^*$, assuming that the constraint:

$$|| x_i(t_i^*) - 2x_i(0) || < || x_i(0) || , \quad i = 1,2,3,... \tag{22}$$

is satisfied. If (22) is not satisfied for some value of i, say I, cell division will not take place if i ≥ I, and the cell stops dividing.

Steady state behaviour in the cell is said to have taken place if there exists a scalar $t_s^* > 0$) such that:

$$\begin{gathered} \lim_{i \to \infty} t_i^* = t_s^* \\ \lim_{i \to \infty} \| \mathrm{x}_{i+1}(t) - \mathrm{x}_i(t) \| = 0, \quad \forall t \in [0, t_s^*] \end{gathered} \tag{23}$$

t_s^* is called the steady-state division time of the cell in this case.

It is assumed in the paper that condition (23) must be satisfied for a normal growing cell or for a cell which survives the effect of an input disturbance applied to it. It is also assumed, for the sake of simplicity, that a cell which stops dividing, cannot survive, i.e. if (22) is not satisfied for all i, then the cell dies.

In [1], rate constants k, initial conditions x(o), inputs u and a division time t^* were found so that (23) is satisfied, producing what is then called a model of a normal cell. The division time of the normal cell obtained was 5.25 sec.

3. ABNORMAL CELL RESULTS

In [1], it was found that if some parameters of the cell are disturbed, the cell may suddenly change into a 'tumor-like' behaviour with a much faster division time and a larger physical size. For example, if $x_8(0)$ is changed from its normal value, the following result occurs (see Table 2):

Table 2 Behaviour of Cell for Change in $x_8(0)$ from its Normal Value

Disturbance Factor	x 7	x 6.5	x 6.0	x 5.5	x 5.0	...	x 1.0
Resultant Steady-state Division time	1.5	1.5	1.65	*4.95	4.70	...	5.3

*Sudden change occurs at this point.

In order to understand how this sudden discontinuity occurs, it is interesting to note how J(t) varies for the case when $x_8(0) = 5.5\ x_8^*(0)$, $x_8(0) = 6.0\ x_8^*(0)$ where $x_8^*(0)$ is the 'normal' parameter value associated with the cell. (See Figures 3a, 3b).

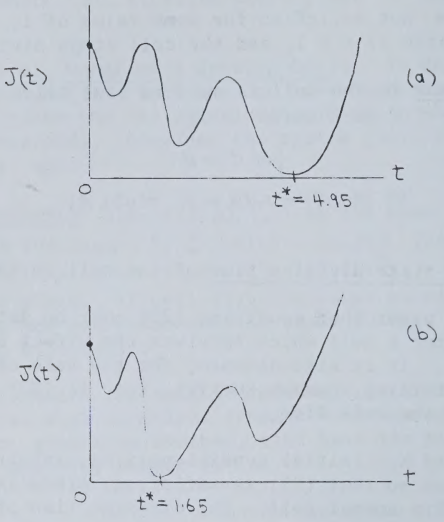


Figure 3 Plot of J(t) vs t for case of disturbance in $x_8(0)$.

(a) Case when $x_8(0) = 5.5 \times x_8^*(0)$; (b) Case when $x_8(0) = 6.0 \times x_8^*(0)$

In this case, it is seen that J(t) has a number of local minima, and depending on the relative values of the local minima, the time at which the global minimum occurs may suddenly shift (in this case from $t^* = 4.95$ to $t^* = 1.65$). This phenomena is suggestive that the "catastrophe theory" of Thom [3], [4] may play a role in abnormal cell behaviour; however, since this theory is basically only concerned with dissipative systems whose state trajectories minimize a potential function, and since J(t) is not a potential for the cell, the possible connection is only suggestive.

3.1 Sensitivity of Tumor-like Cells

The following section gives some sample results obtained on the effect of parameter changes on a 'tumor-like'cell obtained in [1] for the case when K_1 = 5x (normal value of K_1). The steady-state division of time of this tumor-like cell is 0.95.

Table 3 A Study of the Resultant Steady-state Division Time Obtained for a Tumor-like Cell Subject to Parameter Changes.

Parameter	x 10	x 0.1
K_2	1.65 (1.65)	1.05 (1.55)
K_3	dies	1.85 (1.20)
K_4	0.95 (0.55)	1.25 (0.85)
K_5	1.45 (1.00)	0.85 (0.90)
K_6	0.50	1.30 (0.60)
K_7	0.50 (0.60)	1.25 (0.55)
K_8	1.00 (0.70)	0.65 (0.65)
K_9	0.60 (0.70)	1.20 (0.75)

In this case the bracketed qualities of Table 3 correspond to the division time achieved when the disturbed parameter of the cell is changed back to its normal value. It may be seen that in almost all cases the division time of the tumor-like cell remains in a neighbourhood of 0.95. Thus the new tumor-like cell is relatively insensitive to disturbances of its parameters, unlike the case of the normal cell [1]. This observation is also true for most (not all) tumor-like cells obtained in [1].

3.2 Irraditation of Tumor -like Cells

The following section gives some sample results obtained by applying ionizing radiation on the 'tumor-like' cell obtained when K_1 = 5x (normal value of K_1). (See [1] for details of the way this experiment was carried out.)

Table 4 Effect of Ionizing Radiation on the "Tumor-like' Cell Obtained when K_1 = 5x (normal value of K_1)

Radiation Factor	Resulting Steady-state Division Time
r = 1.0 (no radiation)	0.95
r = 0.8 (low radiation)	1.20
r = 0.4	1.80
r = 0.2	2.50
r = 0.1	3.00
r = 0.08	3.45
r = 0.06	4.40
r = 0.052	5.2 (normal cell)
r = 0.04 (high radiation)	cell dies

In this case, it can be seen that as the radiation factor increases, the cell eventually reverts back to its previous normal behaviour. It should be noted however that the above type of radiation effect does not always occur on all types of 'tumor-like' cells.

CONCLUSIONS

This paper has briefly described some properties of the 'tumor-like' cells obtained in [1] when a cell growing in a culture medium is subject to disturbances. In particular, it is shown that the 'tumor-like' cells tend to be less sensitive to changes in the cell's parameters than normal cells. It is also shown that some (not all) 'tumor-like' cells may revert back to their normal behaviour, if they are subject to severe radiation.

TERMINOLOGY AND SYMBOLS

Pools

P_e = Extracellular nutrient pool

P_i = General intracellular metabolic pool

P_a = Amino acid pool for protein synthesis

P_n = Nucleotide pool for RNA synthesis

Enzymes

E = Total protein

E_p = RNA polymerase for messenger RNA (M) synthesis

Genes

G_E = Genes for messenger RNA (M) synthesis

G_p = Genes for messenger RNA (M_p) synthesis

G_B = Genes for the synthesis of RNA fraction of ribosome

G_C = Genes for transport RNA (C) synthesis

Messengers

M = Messenger (RNA) for protein (E) synthesis

M_p = Messenger (RNA) for E_p synthesis

B' = RNA fraction of ribosome

B = Ribosome

C = Transport RNA

N = Ribosome and messenger complex for protein (E) synthesis (template)

N_p = Ribosome and messenger complex for E_p synthesis (template)

Intermediate Complexes

$[CP_a]$ $[E_pB]$

$[E_pP_n]$ $[EP_i]$

$[E_pC]$ X

REFERENCES

1. Davison, E.J. "Simulation of Cell Behaviour: Normal and Abnormal Growth", Bull. of Math. Biology, vol. 37, No. 5, 1975, pp. 427-458.

2. Heinmets F., Analysis of Normal and Abnormal Cell Growth, Plenum Press, New York, 1966.

3. Thom, R. "Topological Models in Biology", Topology, vol. 8, 1969, pp. 313-335.

4. Thom, R. Stability Structurelle et Morphogenese, Benjamin, New York, 1972.

COMPARISON OF DIFFERENT CLASSES OF DYNAMIC MODELS*

R. L. Kashyap†
School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907

ABSTRACT

This paper deals with the Bayesian methods of comparing different types of dynamical structures for representing the given set of observations. Specifically, given that a given process $y(\cdot)$ obeys one of r distinct stochastic difference equations each involving a vector of unknown parameters, we compute the posterior probability that a set of observations $y(1),\ldots,y(N)$ obey the ith equation, after making suitable assumptions about the prior probability distribution of the parameters in each class. The difference equations can be nonlinear in the variable y but should be linear in the parameter vector in it. Once the posterior probability is known, we can find a decision rule to choose between the various structures so as to minimize the average value of a loss function. The optimum decision rule is asymptotically consistent and gives a quantitative explanation for the 'principle of parsimony' often used in the construction of models from empirical data. The decision rule answers a wide variety of questions such as the advisability of a nonlinear transformation of data, the limitations of a model which yields a perfect fit to the data (i.e., zero residual variance) etc. The method can be used not only to compare different types of structures but also to determine a reliable estimate of spectral density of process. We compare the method in detail with the hypothesis testing method, maximum entropy spectral analysis method and other methods and give a number of illustrative examples.

INTRODUCTION

For determining an appropriate representation for a given time series, one usually assumes a certain structure involving a set of parameters and the values of these parameters are estimated from the given data. For many of the geophysical or hydrological time series, the physics of the problem is not well understood to specify a unique structure for the stochastic process. In these cases, one of the main reasons for the construction of models is for understanding the dynamics of the process. Hence instead of restricting ourselves to one particular structure like autoregressive structure, a linear combination of orthogonal polynomials in time t or a fourier series, we should endeavor to quantitatively compare the validity of these widely different structures for representing the given data. For every structure, we should try to determine the probability that the given set of measurements could have come from that structure and choose that structure which has the highest probability.

We will first clarify the notion of structure and model because these words are used in widely differing contexts. Consider a stochastic process $y(\cdot) \varepsilon\ \mathcal{Y}$ which obeys a stochastic difference equation (1.1).

---

*This work was partially supported by Air Force Office of Scientific Research under Grant 74-2661 and National Science Foundation under Grant ENG 74-17586.

†School of Electrical Engineering, Purdue University, W. Lafayette, Indiana 47907.

$$f_i(y(t)) = g_i(t, y(t-1), \ldots, y(t-m_i), \underset{\sim}{\theta}) + w(t), \qquad (1.1)$$

where f_i etc. obey the following assumptions (A1)-(A3).

(A1) $f_i(y)$ is any differentiable function of y such as $\ln y + C$, $y + C$, $y^2 + C$, etc.; f_i: $y \to R$.

The constant C is chosen so that the empirical mean of $f_i(y)$ with the given set of observations $\xi(N) = (y(1), \ldots, y(N))$ is zero, i.e., $C = (1/N) \sum_{t=1}^{N} f_i(y(t))$

(A2) Let $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_{n_i})^T \varepsilon R^{n_i}$. g_i is any function linear in $\underset{\sim}{\theta}$, but not necessarily linear in $y(t-1)$, t etc. m_i is any integer greater than or equal to zero.

g_i: $\tau \times \mathcal{Y}^{m_i} \times R^{n_i} \to R$.

$\tau = \{1, 2, 3, \ldots\}$.

when $m_i = 0$, the function g_i is defined g_i: $\tau \times R^{n_i} \to R$.

The parameters m_i and n_i characterizing the function g_i will be of particular interest later in developing decision rules.

(A3) $\{w(\cdot)\}$ is a sequence of I.I.D. variables with distribution $N(0,\rho)$ and $w(t)$ is independent of $y(t-j)$, $j \geq 1$, $0 \leq \rho < \infty$.

A model is a 4 tuple $\{f_i, g_i, \overline{\theta}, \rho\}$ uniquely associated with the difference eq. (1.1). A process $y(\cdot)$ is said to obey a model $\{f_i, g_i, \underset{\sim}{\theta}, \rho\}$ if it obeys the associated difference equation (1.1).

Let C_i be a class of models having the same functions f and g but differing in $\underset{\sim}{\theta}$ or ρ.

$$C_i \triangleq \{f_i, g_i, m_i, n_i, \textcircled{H}_i\}$$
$$= (f_i, g_i, \underset{\sim}{\theta}, \rho): \ \underset{\sim}{\theta} \ \varepsilon \ {}_i, \ 0 \leq \rho < \infty\}, \qquad (1.2)$$

where

$$\textcircled{H}_i = \{\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_{n_i})^T, \ \theta_i \neq 0 \quad \forall_i, \ \theta \ \varepsilon \ R^{n_i}\} \qquad (1.3)$$

C_i is labeled as the ith class or ith structure. Two classes C_i and C_j are said to be mutually exclusive if they differ in the functions f or g, in a nontrivial way i.e., there is no process which obeys a model in C_i as well as a model in C_j.

Given r mutually exclusive classes or structures $C_1, \ldots, C_r$ and a set of observations $\xi = \{y(1), \ldots, y(N)\}$ our intention is to develop a decision rule to assign the observation set to one of the classes among C_i, $i = 1, \ldots, r$ so as to minimize a suitable criterion function and also determine the probability of error associated with the decision. The loss function csn be chosen to reflect the particular needs of the problem, i.e. If we are investigating class selection solely to obtain a good model for forecasting or for estimating spectral density, then the loss function can be chosen.

We recall that only restriction on the classes C_i is that the function g_i must be linear in $\underset{\sim}{\theta}$. Hence the theory allows us to simultaneously compare widely different structures or classes such as (i) the class of autoregressive (AR) models of order m_1 in y, (ii) the class of AR models of order m_2 in $\ln y$, (iii) the class of models made of m_3 orthogonal polynomials in the variable t, (iv) class of harmonic models involving certain frequencies, using the same data.

Problems of this type are referred to as compound hypothesis testing in the statistical literature [2] and there is no general theory to handle them. There are specific tests for comparing specific pairs of classes such as classes of autoregressive models of 2 different orders. Even here the solution is unsatisfactory for a number of reasons such as (i) the choice of arbitrary quantities like significance levels (ii) lack of any measure of the type II error involved, (the chosen significance level places an upper-bound on the probability of type I error) (iii) the intransitivity of the decision rule does and its lack of any optimality properties etc. There are also many other adhoc rules such as likelihood ratio rule [2] or MAIAC [3] which do not usually possess asymptotic consistency or give a measure of the probability of error of the decision.

Some common problems occurring in the analysis of empirical time series are really problems regarding the appropriateness of the different types of structure for explaining the data. It is being increasingly realized that the traditional methods of spectral analysis involving the use of window functions like those of Bartlett or Hamming [6] often lead to misleading inferences. To overcome such objections, geophysicists use a special method of spectral analysis known as maximum entropy spectral analysis (MESA) whose inferences have greater reliability than those of the traditional methods. The method effectively assumes an autoregressive (AR) structure for the process. If we pose the problem as one of comparing different classes of models having different orders of AR models as done here, we will obtain a solution which overcomes many of the disadvantages of the MESA method, without sacrificing the advantages. Moreover, we can choose the decision rule to minimize a loss function which explicitly reflects the needs in the context namely minimizing the errors in the estimation of spectral density of the process. Similar problems are faced in the spectral analysis as applied to the processing of acoustic signals also [7].

In our paper we will first compute $P(C_i|\xi)$, the posterior probability of the given data having been generated by some model in C_i, for all $i = 1,\ldots,r$, assuming a suitable prior distribution for the classes and the various parameters. Subsequently we will derive optimal decision rules according to various types of criteria and discuss the asymptotic consistency of the decision rule. We will give many types of examples such as (i) whether a nonlinear transformation of the data will yield a better representation of the data, (ii) when is a polynomial fit of the data (or a fourier series representation of the data) with zero residual variance inferior to a relatively small order autoregressive representation etc., to bring out the superiority of the method to existing methods of comparison like hypothesis testing [1,2], maximum entropy spectral analysis [4,5,10], etc.

Even if all the classes have the same prior probability, the posterior probability of class C_i being the correct class for the observation set $\xi(N)$ having N observations is, under appropriate assumptions

$$P(C_i|\xi(N)) = \frac{\exp[0.5\, h_i(\xi(N))]}{\sum_{k=1}^{r} \exp[0.5\, h_k(\xi(N))]}, \quad i = 1,\ldots,r \tag{1.4}$$

$$\begin{aligned} h_i(\xi(N)) = {} & 2N\,\overline{\ln f_i'} - (N-m_i)\ln(\rho_i + (\rho_{fi}-\rho_i)/N) - n_i \ln N \\ & - m_i \ln \rho_{fi} + m_i - (1/\rho_{fi}) \sum_{t=1}^{m_i} (f_i(t))^2 + O(1/N) \qquad i = 1,2,\ldots,r \end{aligned} \tag{1.5}$$

where $\overline{\ln f_i'}$ is the empirical mean of the function $\ln[d\, f_i(y)/dy|y = y(t)]$, ρ_{fi} is the empirical variance of $f_i(y(t))$ and ρ_i is the residual variance. The expression (1.5) clearly brings out the adverse effects of dealing with classes with large n_i, i.e., having too many parameters to be estimated. The role played by the lag variable m_i and the number of estimated parameters

n_i is quite different in general even though in the widely discussed case of autoregressive processes $m_i = n_i$. Thirdly, even if $\rho_i = 0$, as in polynomial fitting models, $P(C_i|\xi(N))$ is still finite. These and other features will be discussed extensively in the subsequent sections.

II. ASSUMPTIONS

For simplicity we will denote a class C_i by the 4 tuple $[f_i, g_i, m_i, n_i]$. The suppressed term ${}_i$ will be mentioned wherever necessary. Further $f_i(t) \triangleq f_i(y(t))$. Similarly the function g_i will be written as $g_i(t,\underset{\sim}{\theta})$ suppressing the other arguments. Moreover g_i can be written as:

$$g_i(t,\underset{\sim}{\theta}) = (z_i(t-1)^T\underset{\sim}{\theta}, \text{ where } z_i \text{ is independent of } \underset{\sim}{\theta},$$

$$\underset{\sim}{z}_i^T(t-1) = \nabla_{\underset{\sim}{\theta}} g_i(t,\underset{\sim}{\theta}),\ n_i \text{ - vector}$$

We consider r classes C_i, $i = 1,\ldots,r$, $C_i = [f_i, g_i, m_i, n_i]$ where f_i and g_i obey the conditions (A1) - (A3). We will make the following additional assumptions on the functions f, g etc.

(A4) The functions f_i and f_j and g_i and g_j obey at least one of the following conditions for every possible pair (i,j), $i \neq j$, $i,j = 1,\ldots,r$
(i) $f_i(y) \neq kf_j(y) + C$ for any k,C and almost all y
(ii) For every $\underset{\sim}{\theta} \in {}_i$, there does not exist a $\underset{\sim}{\theta}' \in {}_j$ satisfying the following relation
$$g_i(t,y(t-1),\ldots,y(t-m_i),\underset{\sim}{\theta}) = g_j(t,y(t-1),\ldots,y(t-m_j),\underset{\sim}{\theta}')$$

(A5) If a process $y(\cdot)$ belongs to one of the classes C_i, $i = 1,2,\ldots,r$, then $[1/N \sum_{t=m_i+1}^{N} \underset{\sim}{z}_i(t-1)z_i^T(t-1)]$ is finite and positive definite for all N and $i = 1,\ldots,r$.

We need the following assumptions regarding the prior probability distribution of θ and ρ and the probability distribution of the initial conditions for eq. (1.1).

(A6) Let $x(t) = f_i(y(t))$. The m_i initial values $y(1),\ldots,y(m_i)$ of a process $y(\cdot)$ obeying the model in C_i obeys the following normal density $p(x(1),\ldots,x(m_i)) \sim N[\underset{\sim}{0},\underset{\sim}{R}_i]$ where $\underset{\sim}{0}$ is the null vector of dimension m_i and $\underset{\sim}{R}_i$ is a $m_i \times m_i$ covariance matrix.

(A7) If $y(\cdot)$ obeys a model in C_i involving m_i lagged variables, then the prior density of $\underset{\sim}{\theta}$ and $\underset{\sim}{\rho}$ obeys the following relation:

$$p(\underset{\sim}{\theta},\rho|y(1),\ldots,y(m_i),C_i) = p(\underset{\sim}{\theta},\rho|C_i)$$

(A8) The variable ρ has the following probability density: $p(\rho|C_i) = \alpha/\rho$, $\alpha > 0$, valid for all i.

(A9) The variable $\underset{\sim}{\theta} \in {}_i$ has the following conditional prior density given ρ:

$$p(\underset{\sim}{\theta}|\rho,C_i) \sim N(\underset{\sim}{\theta}_{0i},\underset{\sim}{S}_{0i}\rho)$$

where $\underset{\sim}{S}_{0i}$ is a

(A10) The prior probability of class C_i is $P(C_i)$, $i = 1,2,\ldots,r$, $0 < P(C_i) < 1$, $\Sigma_{i=1}^r P(C_i) = 1$.

We will discuss the assumptions. The assumption (A4) and condition ((1.3) are sufficient for the classes $C_1,\ldots,C_r$ to be mutually exclusive, i.e. there does not exist a process $y(\cdot)$ which obeys 2 different models in 2 different classes. (A5) is needed for the existence of the matrices occurring in the optimal decision function. To understand (A6) and (A7), we should note that we have not made any assumption about the stationarity of the process y obeying any model in C_i for any i. Specifically, if $y(\cdot)$ belongs to C_i, then $\{y(1),\ldots,y(N)\}$, the observation set is generated as follows: $y(1),\ldots,y(m_i)$ act as initial conditions. Based on $y(1),\ldots,y(m_i)$ and $w(m_i+1)$, $y(m_i+1)$ is generated from (1.1), and $y(t)$, $t > m_i+1$ are generated recursively using (1.1). Clearly the initial conditions $y(1),\ldots,y(m_i)$ cannot throw any light on the parameters $\underset{\sim}{\theta}$ and ρ which characterize that particular model in C_i obeyed by $y(\cdot)$. This statement is the assumption (A7).

The assumption (A3) yields the conditional probability density $p(y(m_i+1),\ldots,y(N)|y(1),\ldots,y(m_i)\underset{\sim}{\theta},\rho,C_i)$ as shown below. This expression in conjunction with (A6) and (A7) yields the joing probability density of all the observations namely $p(y(1),\ldots,y(N)|\underset{\sim}{\theta},\rho,C)$. The details are indicated below.

$$p(x_i(m_i+1),\ldots,x_i(N)|x_i(1),\ldots,x_i(m_i),\underset{\sim}{\theta},\rho,C_i)$$

$$= \prod_{t=m_i+1}^{N_i} p(x_i(t)|x_i(t-1),\ldots,x_i(1),\theta,\rho,C), \text{ by (A3)}$$

$$= \prod_{t=m_i+1}^{N_i} \frac{1}{\sqrt{2\pi\rho}} \exp[-1/2\ (x_i(t) - g_i(t,\theta))^2], \text{ by (A3)} \tag{2.1}$$

Transforming the variables x into y by the relation $x_i(t) = f_i(y(t))$, (2.1) yields:

$$p(y(m_i+1),\ldots,y(N)|y(1),\ldots,y(m_i),\underset{\sim}{\theta},\rho,C_i)$$

$$= \prod_{t=m_i+1}^{N} (f_i'(t)/\sqrt{2\pi\rho})\exp[-1/2\rho(f_i(y(t)) - g_i(t,\underset{\sim}{\theta}))^2], \tag{2.2}$$

where $f_i'(t) = (d\ f_i(y)/dy)|y = y(t)$

$$p(y(1),\ldots,y(m_i)|\theta,\rho,C_i) = p(y(1),\ldots,y(m_i)|C_i), \text{ by (A7)}$$

$$= \frac{\prod_{t=1}^{m_i} f_i'(t)}{(2\pi)^{m_i/2}\ (\det R_i)^{1/2}} \exp[-(1/2)||(f_i(1),\ldots,f_i(m_i)||^2_{R_i^{-1}}], \text{ by (A6)} \tag{2.3}$$

Multiplying the expressions (2.2) and (2.3) yields the required density $p(y(1),\ldots,y(N)|\underset{\sim}{\theta},\rho,C_i)$. The prior density of $\underset{\sim}{\theta}$ given ρ in (A9) is so chosen that the posterior density $p(\theta|\rho,\xi,C_i)$ is also Gaussian. There is an extensive literature on the choice of the prior density of ρ. The density $p(\rho)$ given in (A8) is commonly used in statistical literature and can be defended on many different grounds [8]. However it is called improper since $\int_{a_1}^{\infty} p(\rho)d\rho$ exists only for $a_1 > 0$. But in computing the posterior density, we can allow the limit a_1 to be zero.

We will offer some suggestions for the choice of $\underset{\sim}{\theta}_{0i}$ and $\underset{\sim}{S}_{0i}$, occurring in the prior density $p(\underset{\sim}{\theta}|\rho,C_i)$. The only guideline available for the choice of $\underset{\sim}{S}_{0i}$ is that it be relatively large in view of the great initial uncertainty regarding the value of θ appropriate for the given data. Even though the effect of $\underset{\sim}{\theta}_{0i}$ and $\underset{\sim}{S}_{0i}$ are asymptotically insignificant on the optimal decision, still the arbitrary choice of θ_{0i} and S_{0i} for various i may make the computation of the $p(C_i|\xi)$ more cumbersome than it need be. Accordingly we propose that the following choice for S_{0i}, $i = 1,\ldots,r$.

(A11) $$\underset{\sim}{S}_{0i} = [1/N-m_i \sum_{t=m_i+1}^{N} \underset{\sim}{z}_i(t-1)\underset{\sim}{z}_i^T(t-1)]^{-1}$$

The choice in (A11) is unconventional in the Bayesian literature since it depends on the observations. However, we will show later that such a choice implies the following expression for the posterior variance of $\underset{\sim}{\theta}$ given ξ and ρ

$$\mathrm{Var}[\theta|\xi,\rho,C_i] \triangleq \underset{\sim}{S}_i = \underset{\sim}{S}_{0i}/(N-m_i+1) \tag{2.4}$$

our choice is reasonable since $\underset{\sim}{S}_{0i}\rho$ is much larger than the posterior variance $S_i\rho$. Any other choice for the $\underset{\sim}{S}_{0i}$ would have made the expression for S_i more complicated than the one in (2.4).

Next let us turn our attention to the choice of the vector $\underset{\sim}{\theta}_{i0}$ occurring in $p(\theta|\rho,C_i)$. The obvious choice for $\underset{\sim}{\theta}_{i0}$ is the null vector, stated in (A12).

(A12) $\underset{\sim}{\theta}_{i0} = [0,\ldots,0]^T$, ($n_i$ - vector)

It is important to realize that (A12) is valid only if $f_i(\cdot)$ is chosen as stated in (A1) i.e., $1/N \sum_{t=1}^{N} f_i(y(t)) = 0$. Otherwise, (A12) will be inconsistent with the fact that mean of $w(\cdot)$ is zero in eq. (1.1).

Finally consider the covariance matrix $\underset{\sim}{R}_i$ occurring in (A6) i.e., the probability density of the initial conditions $y(1),\ldots,y(m_i)$. The effect of this term on the final decision rule is not strong. Hence to simplify the computation, we make the assumption (A13).

(A13) $\underset{\sim}{R}_i = \rho_{fi} I$

where ρ_{fi} = empirical variance of $f_i(y(t))$.

III. THE POSTERIOR PROBABILITY $P(C_i|\xi)$

Let $\xi = \{y(1),\ldots,y(N)\}$

$$P(C_i|\xi) = p(\xi|C_i)\, P(C_i)/p(\xi)$$

$$= \int_0^\infty d\rho \int_{\theta\varepsilon} d|\theta|\; p(\xi|\underset{\sim}{\theta},\rho,C_i)p(\underset{\sim}{\theta},\rho|C_i)P(C_i)/p(\xi) \tag{3.1}$$

The expression for $p(\tilde{\xi}|\underset{\sim}{\theta},\rho,C_i)$ has been derived in section II. $p(\theta,\rho|C_i)$ is available from assumptions (A8) and (A9). Hence the integration in (3.1) can be performed as indicated in the appendix 1 leading to the following theorem 1.

Theorem 1: Under the assumptions (A1)-(A3), the posterior probability $P(C_i|\xi)$ has the following form

$$P[C_i|\xi] = K \exp[0.5\, h_i(\xi)]$$

where

$$K = 1/\sum_{i=1}^{r} \exp[0.5\, h_i(\xi)]$$

$$h_i(\xi) = 2N\,\overline{\ln f_i'} - (N-m_i)\ln(\rho_i + (\rho_{fi}-\rho_i)/N)$$
$$+ 2\ln P(C_i) - n_i \ln N - m_i \ln \rho_{fi} + G_1(m_i) + O(1/N)$$

$$G_1(m_i) = m_i - (1/\rho_{fi}) \sum_{t=1}^{m_i} (f_i(t))^2$$

$$\rho_i = 1/(N-m_i) \sum_{t=m_i+1}^{N} (f_i(t) - \underset{\sim}{z}_i^T(t-1)\underset{\sim}{\theta}^*)^2$$

$$\underset{\sim}{\theta}^* = [(N-m_i+1)/(N-m_i)][\sum_{t=m_i+1}^{N} z_i(t-1)z_i(t-1)]^{-1} \sum_{t=m_i+1}^{N} z_i(t-1)f_i(t)$$

$$\overline{\ln f_i'} = (1/N)\sum_{t=1}^{N} \ln f_i'(t),\quad f_i'(t) = d\, f_i(y)/dy\Big|_{y=y(t)}$$

$$\rho_{fi} = (1/N)\sum_{t=1}^{N} (f_i(y(t))^2 \triangleq \text{empirical variance of } f_i(y(t),\ldots,f_i(y(N))$$

A proof of Theorem 1 is in appendix 1.

Comment 1: An expression for the $P(C_i|\xi(N))$ without the assumptions (A11)-(A13) regarding the parameters of prior distribution is given in lemma 1 in appendix 1. Obviously it is computationally more complicated.

Comment 2: $E[G_1(m_i)|C_i] = 0$
Variance $[G_1(m_i)|C_i] = 2m_i$

Hence $G_1(m_i)$ is of the order $O(\sqrt{m_i})$. While comparing 2 classes having different values of m_i, the term $G_1(m_i)$ does not make much difference in comparison with other terms and hence can be neglected when N is large.

Comment 3: The model $(f_i,g_i,\underset{\sim}{\theta}_i^*,\rho_i^*)$ is the best fitting model in the class C_i for the given data ξ. Alternatively, if ξ obeys some (unknown) model $(f_i,g_i,\underset{\sim}{\theta},\rho)$ in C_i, then θ_i^* and ρ_i^* are the Bayesian estimates of $\underset{\sim}{\theta}$ and ρ.

We will discuss many other features of the posterior probability function $P(C_i|\xi)$ in section V.

IV. OPTIMAL DECISION RULE

1. Optimal Decision Rules:

Let $d(\xi(N))$ be a decision rule where d is a map:
$d: \mathcal{Y}^N \rightarrow \{C_1,\ldots,C_n\}$

Consider a loss function L which reflects the cost of wrong assignment of the observation set ξ to a class C_j using the rule.

$$L(C_i,d(\xi) = C_j) = 0 \text{ if } C_i = C_j$$
$$= w_{ij} > 0, \text{ if } C_i \neq C_j$$

Our intention is to choose the decision rule to minimize the average value of the loss function L, i.e., minimize J(d)

$$J(d) = E[L(C_i,d(\xi))]$$
$$= \sum_{i=1}^{r} P(C_i)\int L(C_i,d(\xi))P(\xi|C_i)d|\xi|$$

or

$$J(d(\xi) = C_j) = \int (\sum_{i=1}^{r} w_{ij}P(C_i|\xi))p(\xi)\ d|\xi|$$

The optimum decision rule is

$$d^*(\xi) = C_k = \underset{C_j\varepsilon\{C_1,\ldots,C_r\}}{\text{Argument Minimum}} \sum_{i=1}^{r} w_{ij}P(C_i|\xi) \tag{4.1}$$

The loss function or, in particular, the weights w_{ij} can be chosen to reflect the particular needs of the problem. If we are interested in minimizing the probability of error in the assignment of class to ξ, the choice of w_{ij} is

$$w_{ij} = 0 \text{ if } i = j$$
$$= 1 \text{ if } i \neq j$$

In that case, the optimum decision rule is;

$$d^*(\xi) = C_k = \text{Argument}[\underset{C_j\varepsilon\{C_1,\ldots,C_r\}}{\text{Maximum}} P(C_j|\xi)] \tag{4.2}$$

i.e. The decision rule assigns ξ to the class having the highest posterior probability.

Probability of error of the optimal decision rule (4.2) =

$$[1-\underset{C_j}{\text{Max}}\ P(C_j|\xi)] \tag{4.3}$$

On the other hand, if our sole interest in class selection is to obtain a reliable estimate of spectral density, then we should choose w_{ij} as follows

$$w_{ij} = \int_{-\omega_N}^{\omega_N} d\omega (\ln S_i(\omega) - \ln S_j(\omega))^2$$

where $S_i(\omega)$ is the "best" estimate of the spectral density of the process based on ξ given that the process ξ belongs to class C_i. In that case, the optimum decision rule (4.1) simplifies

$$d^*(\xi) = C_j \text{ if } (\ln P(C_j|\xi)-H)^2 \leq (\ln P(C_i|\xi)-H^2 \quad \forall\ i \neq j, \tag{4.4}$$

with

$$H = \sum_{i=1}^{r} P(C_i|\xi) \ln P(C_i|\xi) \tag{4.5}$$

The 2 illustrations should be sufficient to reveal the power of the decision rule to reflect the needs of the particular problem.

V. DISCUSSION AND COMPARISON

We will highlight some of the important features of our decision rules. Most of these features are absent in the decision rules based on hypothesis testing and other methods which will be discussed subsequently.

(P1) The decision rule can compare simultaneously many classes obeying the conditions in Sections I and II.

(P2) The decision rule is transitive, i.e. Let $C_i \quad C_j$ mean that in comparing the classes C_i and C_j, the decision rule prefers C_i to C_j. Then $C_1 \quad C_2$ and $C_2 \quad C_3 \Rightarrow C_1 \quad C_3$ provided all the classes are equiprobable.

(P3) The decision rule is asymptotically consistent, i.e., while comparing 2 classes C_1 and C_2 based on the observation set $\xi(N)$, then $\lim_{N\to\infty} P(C_2|\xi(N))/P(C_1|\xi(N)) \to \infty$ if the observation set comes from some (unknown) member in C_2 and the error probability $P(C_1|\xi(N))$ decays at least as fast as $N^{-(n_1-n_2)/2}$ if $n_1 > n_2$ or k_2^{-N} where $k_2 > 1$.

(P4) The decision rule is optimal in the sense that it minimizes a suitable loss function which reflects the needs of the particular problem.

(P5) An explicit expression is available for the probability of error in the decision given by the rule.

(P6) The only arbitrary quantities appearing in the decision rule are the prior probabilities of the classes. There are no other arbitrary quantities like significance levels. The effect of the assumption about the prior probabilities is asymptotically negligible unlike the significance levels used in the hypothesis testing methods. Typically the prior probabilities of the classes can be assumed to be equal.

(P7) The roles played by m_i and n_i in the decision rules are quite different. The contribution of terms involving m_i to $\ln P(C_i|\xi)$ is $O(1)$ where as the contribution of terms involving n_i is $O(\ln N)$.

(P8) The posterior probability $P(C_i|\xi)$ involves a term $\exp[-n_i \ln N]$ even if all the r classes are a priori, equiprobable. When we are comparing 2 classes C_i, $i = 1,2$ such that they yield the same value of ρ_1 and ρ_2 and $\ln f_i^!$, $i = 1,2$, the posterior probability of the class having smaller n_i will tend to 1 as N tends to infinity. This is a quantitative proof of the "principle of parsimony" which states that if 2 models explain the same data (i.e., have same ρ), the model involving the smaller number of estimated parameters has a higher plausibility of being the correct model of the process than the other model.

(P9) The decision rule for comparing classes is valid even if one of the members of a class yields a zero resicual variance (i.e., $\rho_i = 0$) with the given data. This is possible because ρ always appears in the decision rule in the form $\ln[\rho + (\rho_f - \rho)/N]$. Hence even if ρ is zero or very small, $\ln[\rho + (\rho_f - \rho)/N]$ is still finite. This property is very useful in comparing stochastic models like autoregressive models involving a small number of parameters with polynomial or fourier series models which involve a large number of parameters, but yield low residual variance.

Further details and illustrative edamples are contained in the following technical report which available from the author.

R. L. Kashyap, "A Bayesian Comparison of Different Classes of Dynamic Models Using Empirical Data", Tech. Report TR-EE 76-40, School of Electrical Engineering, Purdue University.

VI. REFERENCES

[1] R. L. Kashyap and A. R. Rao, Dynamic Stochastic Models from Empirical Data, Academic Press, 1976.

[2] E. L. Lehmann, Testing Statistical Hypothesis, Wiley, New York, 1959.

[3] H. Skaike, "A New Look at Statistical Model Identification", IEEE Trans. on Automatic Control, Vol. AC-19, Dec. 1974, p. 716-722.

[4] D. E. Smylie, G. K. C. Clarke and T. J. Ulrych, "Analysis of Irregularities in the Earth's Rotation" in "Methods in Computational Physics", vol. 13, Geophysics, (ed.) Alder et al.

[5] J. P. Burg, "The Relationship Between Maximum Entropy Spectra and Maximum Likelihood Spectra", Geophysics, vol. 37, p. 375-376.

[6] M. S. Bartlett, An Introduction to Stochastic Processes, Cambridge Univ. Press, London, 1966.

[7] J. D. Markel, "Digital Inverse Filtering: A New Tool for Formant Trajectory Estimation", IEEE Trans. Audio Electroacoustics, AU-20, 1972, p. 129-137.

[8] Kashyap, R. L., "Probability and Uncertainty," IEEE Trans. Information Theory, 1971, pp. 641-650.

[9] G. U. Yule, "On the Method of Investigating Periodicities in Disturbed Series with Special Reference to Wolfer's Sunspot Numbers", Phil. Trans. A226, 1927, p. 267.

[10] N. Andersen, "On the Calculation of Filter Coefficients for Maximum Entropy Spectral Analysis", Geophysics, vol. 49, no. 1, Feb. 1974, pp. 69-72.

LINEAR PREDICTIVE CODING WITH ADDITIVE NOISE FOR APPLICATION TO SPEECH DIGITIZATION

JAMES L. MELSA AND JAMES D. TOMCIK
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana 46556

ABSTRACT

There is a great deal of current interest in interconnecting waveform coding processors for speech signals, such as delta modulators, with linear predictive coders. This type of interconnection causes the input speech signal to the linear predictive coding system to be corrupted by quantization noise. As a result, inaccuracies in the identified predictor parameters result. This paper presents an evaluation of two different strategies designed to reduce the effect of input quantization noise on such systems.

1. INTRODUCTION

At the present time there are two major approaches to the speech digitization problem. Wide-band encoding methods such as Delta Modulation attempt to encode the speech waveform itself, while narrow-band Linear Predictive Coders (LPC) attempt to model the vocal tract and its excitation mechanism.

Since both approaches are being used in various applications, it has become desirable to design an interface so that communication between terminals utilizing waveform encoders and terminals using Predictive Coders is possible. Thus, the interface problem is really a two-fold problem: that of LPC to waveform encoding and waveform encoding to LPC.

In this paper, the problem of interfacing waveform encoders to Linear Predictive Coders is studied. One manner of achieving an interface is to reconstruct the speech using a waveform reconstructor, and then to analyze the reconstructed waveform with an LPC analyzer. This method is illustrated in Figure 1.

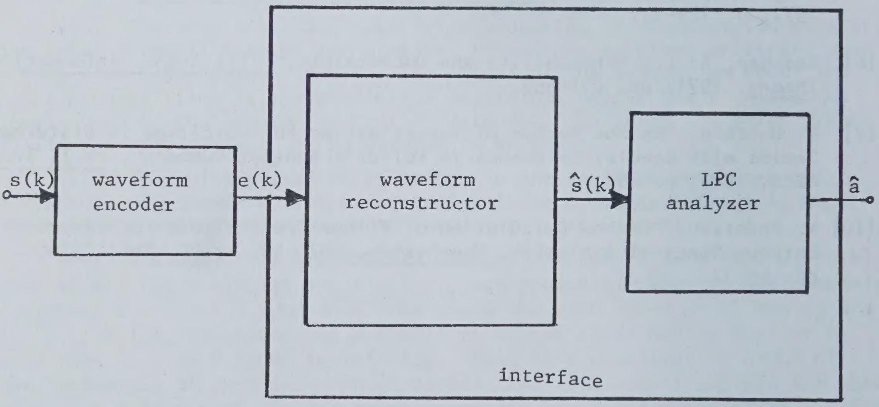


Fig. 1 A Natural Interface Between Wide-Band and Narrow-Band Speech Environments

In the illustration, $s(k)$ is the original sampled speech, $e(k)$ is the transmitted error signal, $\hat{s}(k)$ is a reconstructed version of $s(k)$ and $\hat{a}$ is a vector of LPC feature parameters. The signal $\hat{s}(k)$, unfortunately, is corrupted by quantization noise due to the quantization of $e(k)$. As a result of this noise, severe errors can occur in the identified feature parameters $\hat{a}$. Thus, the interface problem can be viewed as the problem of estimating LPC parameters from speech corrupted by random noise. This is the approach taken in this paper.

Section 2 will present a formulation of the problem. A modified Atal-Hanauer [4] Algorithm is presented in Section 3, and an extended Kalman Filter Formulation is described in Section 4. Section 5 presents the conclusions based on this study.

2. PROBLEM FORMULATION

The sounds of speech are of essentially two types, depending on what sort of excitation is applied to the vocal tract. If a continuous breath is applied, the sound produced resembles white noise and is called unvoiced speech. If the excitation takes the form of pulses, the sound produced is said to be voiced. Thus, it is possible to view the vocal tract as a filter which is excited by two different types of waveforms.

This observation forms the basis for the Linear Predictive Model of speech production. The model was used by Atal and Hanauer [4] in their formulation of the Covariance Method. This model assumes that speech may be modelled as the output of a time varying linear filter described by:

$$s(k) = \sum_{i=1}^{N} a_i s(k-i) + Gw(k) \tag{1}$$

where N is the order of the filter, G is a parameter related to the RMS value of the speech, $s(k)$ is a speech sample at time k and a_i, $i = 1, 2, \ldots, n$, are the LPC feature parameters or predictor coefficients which describe the shaping of the excitation function w by the vocal tract. To model the quantization noise introduced by waveform encoding techniqes, the observed signal is assumed to be:

$$s^*(k) = s(k) + v(k) \tag{2}$$

where $v(k)$ is a zero-mean white noise with variance V_v which is assumed to be constant. Thus, the difficulty involved with estimating predictor coefficients can be seen from (2). If only s^* is available, the recursiveness of the model (1) is lost. One method of solving this problem is to estimate the speech sample $s(k)$ from $s^*(k)$. This is the approach described in Sec. 4. Alternatively, one may perform a standard LPC analysis using $s^*(k)$ instead of $s(k)$. In the next section this is shown to introduce a bias on the resulting predictor coefficients. This bias is correctable, however, to some extent, and methods for effecting this correction will be discussed in detail.

3. BIAS CORRECTION FOR BLOCK ANALYSIS

In the block analysis approach to narrow-band linear predictive coding of speech signals, a frame of M samples is analyzed to obtain estimates of the N predictor coefficients a_i. This method was first proposed by Atal and Hanauer [4]; later Markel [5] presented a related method which used special windowing to achieve computational advantages.

The predictor coefficients are selected to minimize the mean square prediction error over the analysis frame which is given by

$$<e^2> = \frac{1}{M} \sum_{m=0}^{M-1} (s(m) - \sum_{i=1}^{N} a_i s(m-i))^2 \tag{3}$$

By setting the partial derivative of each a_i to zero, the following expression is obtained

$$\sum_{i=1}^{N} a_i [\frac{1}{M} \sum_{m=0}^{M-1} s(m-j)s(m-k)] = \frac{1}{M} \sum_{m=0}^{M-1} s(m)s(m-j) \tag{4}$$

for j = 1, 2,..., N. Let the covariance matrix R be defined by

$$R_{ij} = \frac{1}{M} \sum_{m=0}^{M-1} s(m-j)s(m-i) \tag{5}$$

and the vector ϕ be

$$\phi_j = \frac{1}{M} \sum_{m=0}^{M-1} s(m)s(m-j) \tag{6}$$

then the vector of predictor coefficients is given by

$$Ra = \phi \tag{7}$$

The matrix R is always positive definite and symmetric and will be Toeplitz if the Markel windowing is used. The solution of Eq. (7) yields the desired predictor coefficent estimates.

The use of PARCOR (6) coefficients k_i provides a direct method for filter stabilization, estimation comparison and improved quantization. The PARCOR coefficients are generated from the predictor coefficients a_i by the following algorithm.

$$k_n = b_n^n$$

$$b_i^{n-1} = (b_i^n - k_n b_{n-i}^n)/(1-k_n^2)$$

for i = 1, 2,..., n-1 where initially $b_i^N = -a_i$, 1 = 1,..., N. The algorithm is started with $k_N = b_N^N$ and the coefficients b_i^{N-1}, i = 1, 2,..., N-1 are calculated.

When white noise is added to the speech signal, the elements of the covariance matrix become

$$R_{ij}^* = \frac{1}{M} \sum_{m=0}^{M-1} [s(m-j) + v(m-j)][s(m-i) + v(m-i)] = R_{ij} + \frac{1}{M} \sum_{m=0}^{M-1} v(m-j)v(m-i)$$
$$+ \frac{1}{M} \sum_{m=0}^{M-1} [s(m-i)v(m-j) + s(m-j)v(m-i)] \tag{8}$$

The cross product terms are approximately zero for modest size M because the signal is slowly varying with respect to the noise sequence. If the frame length M is large enough,

$$\frac{1}{M} \sum_{m=0}^{M-1} v(m-j)v(m-i) \sim E\{v(m-j)v(m-i)\} = V_v \delta_K(j-i) \tag{9}$$

so that R^* is

$$R^* = R + V_v I$$

Using the same assumptions, one can show that the ϕ vector is unchanged. The predictor coefficients for the noisy case are given by the solution of

$$R^* a = \phi$$

The solution of this equation will yield biased estimates for a. If we have a reliable estimate of V_v, it can be extracted from R^* to yield unbiased estimates.

Computer simulations of real speech were made to check these results. When R^* was corrected by subtracting $V_v I$, there were some improvements but large errors were also found in some frames. In these frames, the correction caused R to become nonpositive definite. To alleviate this problem, a correction of only $0.8V_v I$ was tried. In this case, only a few frames were bad but some of the improvement in others had been lost also. The next step was to try an adaptive correction of $\delta V_v I$ where $\delta = 1$ initially. If this correction did not produce a stable filter, as determined by checking the PARCOR coefficients, $\delta = \delta - 0.1$ and the process was repeated.

Table 1 shows the result for this study and indicates a significant performance improvement using the adaptive system. The speech for this study is the word "cats" sampled at 8 KHz. The analysis frame length M was 160.

Table 1

Results of Bias Correction Study

k_1 = First PARCOR Coefficient.

Frame No.	No Noise	$V_v = 100$ No Correction	$V_v = 100$ With Correction	Final Value of δ
1	-0.8289	-0.7974	-0.8654	0.9
2	-0.7876	-0.7583	-0.7768	1.0
3	-0.7809	-0.7599	-0.7946	1.0
4	-0.7465	-0.7328	-0.7476	0.7
5	-0.7537	-0.7345	-0.7550	1.0
6	-0.7683	-0.7382	-0.7682	1.0

k_2 = Second PARCOR Coefficient.

Frame No.	No Noise	$V_v = 100$ No Correction	$V_v = 100$ With Correction	
1	0.5556	0.4019	0.7676	
2	0.6312	0.4870	0.5746	
3	0.7160	0.6191	0.7463	
4	0.6838	0.5945	0.6703	
5	0.6619	0.5846	0.67850	
6	0.5201	0.4118	0.5192	

k_3 = Third PARCOR Coefficient

Frame No.	No Noise	$V_v = 100$ No Correction	$V_v = 100$ With Correction	
1	-0.1497	0.0350	0.05736	
2	-0.1764	0.0326	-0.0844	
3	-0.1069	0.0781	-0.0757	
4	-0.1622	0.0432	-0.0896	
5	-0.1083	0.0519	-0.1129	
6	-0.1916	-0.0292	-0.1476	

4. EXTENDED KALMAN FILTER

The given model is as follows:

$$a(k+1) = a(k) + w(k) \tag{12}$$

$$S_N(k+1) = A\,S_n(k) + \Gamma' u(k) \tag{13}$$

$$z(k) = H' S_N(k) + v(k) \tag{14}$$

where $a \in R^N$ is a vector of predictor coefficients assumed constant over an analysis frame, $S_N(k) = [s(k), s(k-1), \ldots, s(k-N+1)]$ is a vector of past speech samples, $w(k)$ is a vector of white state noise having covariance matrix V_w, $u(k)$ is a scalar white noise accounting for modelling errors, and $v(k)$ is white (Gaussian) measurement noise. The matrices above are defined as:

$$A = \begin{bmatrix} a_1 & a_2 & \ldots & a_N \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ 0 & \ldots & 0\ 1 & 0 \end{bmatrix} \qquad H' = [1\ 0 \ldots 0] \qquad \Gamma' = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

A non-linear model for this system can be formulated as:

$$x(k+1) = \Phi(x(k)) + \Gamma\Omega(k) \tag{15}$$

Where:

$$x(k) = \begin{bmatrix} a(k) \\ \hline S_N(k) \end{bmatrix} \qquad \Omega(k) = \begin{bmatrix} w(k) \\ \hline u(k) \end{bmatrix}$$

and

$$\Gamma = \left[\begin{array}{c|c} I_N & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ \hline 0---0 & 1 \\ \hline \begin{matrix} 0------0 \\ \vdots \\ 0------0 \end{matrix} & \end{array}\right] \left.\vphantom{\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}}\right\} N-1 \quad = \begin{bmatrix} I_{N+1} \\ \hline [0]_{N-1 \times N+1} \end{bmatrix}$$

The state transition function $\Phi(\cdot)$ can be expressed as:

$$\Phi[x(k)] = Px(k) + H^T[x^T(k)Qx(k)]$$

Where:

$$P = \left[\begin{array}{c|c|c} I_N & \multicolumn{2}{c}{[0]_{N\times N}} \\ \hline \multicolumn{3}{c}{0------------0} \\ \hline [0]_{N-1\times N} & I_{N-1} & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \end{array}\right] \qquad Q = \tfrac{1}{2}\left[\begin{array}{c|c} [0]_{N\times N} & I_N \\ \hline I_N & [0]_{N\times N} \end{array}\right]$$

and:

$$H = [0 \ldots 0\ 1\ 0 \ldots 0]$$

The observation model can be written as

$$z(k) = Hx(k) + v(k)$$

with H defined as above.

Given this non-linear system model, the extended Kalman Filter Algorithm [2] can be applied directly. The filter equation is:

$$\hat{x}(k+1) = \Phi(\hat{x}(k)) + K(k+1)\ \{z(k+1) - H\Phi(\hat{x}(k))\}$$

Substituting the expression for $\Phi(\cdot)$ and simplifying the result gives

$$\hat{x}(k+1) = P\hat{x}(k) + H^T\ \hat{x}^T(k)Q\hat{x}(k) + K(k+1)\ \{z(k+1) - \hat{x}^T(k)M\hat{x}(k)\}$$

The Kalman gain form is:

$$K(k+1) = V_{\tilde{x}}(k+1)\ \frac{\partial h^T(\hat{x}(k+1 \mid k))}{\partial \hat{x}(k+1 \mid k)}\ V_v^{-1}$$

Where h( ) is the observation function. For this model, the derivative is merely the matrix H^T and:

$$K(k+1) = V_x(k+1)H^T\ V_v^{-1}$$

For the a-priori variance relation, we have:

$$V_{\tilde{x}}(k+1 \mid k) = \frac{\partial \Phi(\hat{x}(k))}{\partial \hat{x}(k)}\ V_{\tilde{x}}(k)\ \frac{\partial \Phi^T(\hat{x}(k))}{\partial \hat{x}(k)} + \Gamma V_\Omega \Gamma$$

Thus:

$$V_{\tilde{x}}(k+1\quad k) = [P+H^T\hat{x}^T(k)2Q]V_{\tilde{x}}(k)[P+H^T\hat{x}^T(k)2Q]^T$$
$$[P+2H^Tx(k)^TQ]V_{\tilde{x}}(k)[P+2H^Tx(k)^TQ]^T + \Gamma V_\Omega \Gamma^T$$

Finally, the a-posteriori error covariance equation is written as:

$$V_{\tilde{x}}(k+1) = V_{\tilde{x}}(k+1 \mid k) - V_{\tilde{x}}(k+1 \mid k)H^T[HV_{\tilde{x}}(k+1 \mid k)H^T+V_v]^{-1}\ HV_{\tilde{x}}(k+1 \mid k)$$

It should be noted that due to the sparse nature of the matrices defined, many simplifications result which make the filter attractive for implementation in a real-time processor.

It has been shown [3] that the choice of $V_\Omega = [0]$, $V_v = 1$, and $V_x(0) = d.I$, where d is a very large number, reduces the Kalman Filter for the non-noisy case to a least-squares estimator. In the noise corrupted case, this does not appear to be the case. In particular, simulations with $V_\Omega \neq [0]$, and $V_v \neq 1$ appear to yield coefficient values nearer to the correct values than simulations using the above parameters.

Thus, an optimization study was undertaken in order to arrive at values for the parameters V_w, V_u and $V_{\tilde{x}}(0)$. For this study, the word "cats," as uttered by a male speaker, was chosen as a representative speech segment. An Atal-Hanauer Analysis was used for a fourth-order model with an analysis frame size of 160 samples. The PARCOR coefficients were then determined and used as the standard against which the PARCOR coefficients, resulting from the non-linear analysis, were compared. Next, since the low order coefficients are the most significant ones in terms of speech quality, the following performance index was postulated:

$$J = \sum_{j=1}^{F_T} \sum_{j=1}^{N} (N-i+1)(\hat{k}_i(j)-k_i(j))^2$$

Where $k_i(j)$ is the i^{th} PARCOR coefficient identified in the j^{th} frame, and $\hat{k}_i(j)$ is the non-linear estimate of that coefficient. The factor (N-i+1) provides a linear weighting to the coefficient errors which is proportional to the significance of the particular coefficient.

The minimization of J yields the following parameter values:

$$V_w = 0$$

$$V_u = 10^6$$

$$V_{\tilde{x}}(0) = 10^{11}$$

Thus, V_u, the variance of the noise used to account for modelling errors must be increased. Table 2 summarizes the performance of the non-linear filter when the above parameters are used.

From the table it can be seen that the coefficients identified from the noisy speech are, in fact, quite close to those identified in the noiseless case. The average magnitude of coefficient error for the modified Atal-Hanauer Algorithm is .0106 for k_1, .0550 for k_2 and .0752 for k_3. Thus, the accuracy of the extended Kalman Filter is better than that of the Atal Algorithm for the high order PARCOR coefficients, and the two are almost equally accurate for low order coefficients.

The reason that the noiseless PARCOR coefficients found in this study differ with those used in Section 3 is that in Section the model order was set at N = 8 and in this study, N = 4 was used. This accuracy, as well as the computational simplicity of the extended Kalman Filter, make this scheme an attractive one for use in identifying noisy speech.

5. CONCLUSIONS

In summary, this research has shown that the natural interface between wide-band and narrow-band speech transmission systems is an adequate solution to the problem, provided that the LPC analyzer used is designed to take into consideration the quantization noise induced into the system by waveform encoding techniques.

While implementation of the modified Atal-Hanauer Algorithm should be an easy modification to existing systems, the modified algorithm is an iterative one and, thus, requires a variable computation time to identify the predictor coefficients. For new or future systems, the accuracy of the extended Kalman Filter Algorithm would seem to be a preferable alternative to the modified Atal-Hanauer Algorithm. A drawback to the Kalman Filter is,

Table 2

Results of Kalman Filter Study

k_1 = First PARCOR Coefficient

Frame No.	No Noise	$V_v = 100$ Kalman Filter	$\lvert \hat{k}_1 - k_1 \rvert$
1	-.7955	-.7767	.0188
2	-.7985	-.7919	.0066
3	-.7802	-.7736	.0066
4	-.7562	-.7756	.0194
5	-.7405	-.7380	.0025
6	-.7712	-.7594	.0118
$\lvert e \rvert_{AV}$	----	----	.0110

k_2 = Second PARCOR Coefficient

Frame No.	No Noise	$V_v = 100$ Kalman Filter	$\lvert \hat{k}_2 - k_2 \rvert$
1	.4953	.4220	.0773
2	.6488	.6184	.0304
3	.7074	.6744	.0330
4	.6990	.6651	.0339
5	.6387	.6139	.0248
6	.5317	.4638	.0679
$\lvert e \rvert_{AV}$	----	----	.0439

k_3 = Third PARCOR Coefficient

Frame No.	No Noise	$V_v = 100$ Kalman Filter	$\lvert \hat{k}_3 - k_3 \rvert$
1	-.2274	-.1757	.0517
2	-.1469	-.0835	.0634
3	-.1191	-.0397	.0794
4	-.1275	-.0778	.0497
5	-.1437	-.0790	.0647
6	-.1856	-.1186	.0674
$\lvert e \rvert_{AV}$	----	----	.0627

of course, its complexity when compared to the block oriented algorithm. The formulation given above, while increasing the model order, does present a savings in terms of the number of multiplies and adds required for implementation

6. ACKNOWLEDGMENT

This work was supported, in part, by the National Science Foundation under Grant ENG 76-00775.

7. REFERENCES

(1) Atal and Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave" J. Acoust. Soc. Am., Vol. 50, pp. 637-655.

(2) A. P. Sage and J. L. Melsa, Estimation Theory with Applications to Communications and Control, McGraw-Hill Company, New York, 1971.

(3) J. Tomcik, "An Application of Kalman Filtering Techniques to Constant Bit Rate Speech Communication," M.S.E.E. Thesis, University of Notre Dame, 1976.

(4) B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am., Vol. 50, Part 2, pp. 637-655, 1971.

(5) J. D. Markel, "Digital Inverse Filtering - A New Tool for Format Trajectory Estimation," IEEE Trans. Aucio Electroacoust., Vol. AU-20, pp. 129-137, June 1972.

(6) F. Itakurg, S. Saito, T. Koike, H. Sawabe, and M. Nishikawa, "An Audio Response Unit Based on Partial Autocorrelation," IEE Trans. Comm., Vol. COM-20, pp. 792-797, August 1972.

IDENTIFICATION PROBLEMS IN ANTICOAGULANT DRUG THERAPY[*]

William F. Powers and Peter H. Abbrecht
University of Michigan
Ann Arbor, Michigan

ABSTRACT

A physically motivated mathematical model with unknown, patient-dependent parameters is employed for the development of anticoagulant drug dosage strategies. An extended Kalman filter identification scheme is employed to determine the parameters, and two optimization problems are solved to determine loading dosages and the optimal measurement time for identification.

1. INTRODUCTION

A number of human disorders require long-term administration of oral anticoagulant drugs. The goal of such drugs is to regulate the blood's prothrombin complex acitivity within relatively narrow limits, with violation of the upper and lower limits resulting in lack of therapeutic effect and possible hemorrhage, respectively.

In Ref. 1 a mathematical model for the quantification of the effects of the anticoagulant drug crystalline sodium warfarin upon the prothrombin complex system was developed and tested on four dogs with excellent agreement between model predictions and experimental results. The model is as follows:

$$\dot{G} = -K_g G + D(t) \tag{1.1}$$

$$\dot{Q} = K_g G - K_e Q \tag{1.2}$$

$$\dot{P} = S_m (1 - F/(F + K_m)) - K_p P \tag{1.3}$$

$$F \equiv 0.03\, Q/\text{warfarin distribution volume} \equiv a\, Q \tag{1.4}$$

Equation (1.1) assumes that the rate of absorption of warfarin from the intestinal tract is proportional to the amount in the intestines, G. The warfarin dosage schedule is described by the function D(t). Equation (1.2) states that the instantaneous rate of change of the amount of warfarin in the body, Q , is given by the difference between its rate of absorption from the gastrointestinal tract, and its rate of metabolism, the latter being con-

---

[*] This research was supported by the National Science Foundation under Grant ENG-75-13641 and a grant from the Michigan Heart Association.

sidered a first order process.

Equation (1.3) states that the rate of change of the prothrombin complex activity (P) is equal to the difference between the rate of synthesis of the complex and its rate of degradation. The reduction in the rate of prothrombin complex synthesis below the normal value (S_m) due to inhibition by warfarin is given by a Michaelis-Menten formulation, in which F is the concentration of free warfarin and K_m the Michaelis constant. While this model is markedly oversimplified in comparison to the actual system, it appears to be a reasonable approach to the problem of obtaining a simple and practical guide for determination of dosage clinically.

Further remarks about this model, other models, and anticoagulant therapy in general may be found in Ref. 2. It should be noted here, however, that current procedures for the determination of a warfarin dosage schedule are rather expensive in that hospitalization is required to get the patient regulated initially. A major goal of this research is to eliminate or reduce in time such hospitalization by employing the mathematical model described above and modern systems techniques.

2. PARAMETER IDENTIFICATION

The mathematical model of Eqs. (1.1) - (1.4) is a nonlinear system with four unknown, patient-dependent parameters K_g, K_e, K_m, and K_p. In the clinical setting, values of the prothrombin complex activity may be determined by the Quick prothrombin time test [3], and such measurements are employed for parameter identification.

Since the state of the system is also desired (in addition to updated parameter values) at each measurement time, the combination state estimation-parameter identification problem is attacked by means of an extended Kalman filter. That is, the system Eqs. (1.1) - (1.4) are augmented by the equations:

$$\dot{K}_g = 0, \dot{K}_e = 0, \dot{K}_p = 0, \dot{K}_m = 0, \qquad (2.1)$$

and the estimation-identification problem becomes a nonlinear filtering problem with continuous dynamics and discrete measurements:

$$x \equiv (G, Q, P, K_g, K_e, K_p, K_m) \qquad (2.2)$$

$$\dot{x} = f[x, D(t)] \qquad (2.3)$$

$$y(t_i) = P(t_i) + v(t_i), \ (i = 0, 1, 2, \ldots) \qquad (2.4)$$

where f denotes the right hand sides of Eqs. (1.1) - (1.3) and (2.1), and $y(t_i)$ is the measurement of P at t_i with noise $v(t_i)$.

For the extended Kalman filter approach to be applicable, one must have reasonable a priori statistics of the initial state and measurement noise. Such statistics exist, and their development and simplification are described in Ref. 2 with the result:

$$\hat{x}\,(t_o) = [\ G_o\ \ 0.0\ \ \ 100.0\ \ \ 6.0\ \ \ 0.60\ \ \ 0.75\ \ \ 0.018\ ] \tag{2.5}$$

$$\text{diag}\ S\,(t_o) = [\ 0.0\ \ \ 0.0\ \ \ 0.0\ \ \ 0.1\ \ \ 1.0\ \ \ 0.1\ \ \ 0.4 \times 10^{-3}\ ] \tag{2.6}$$

$$E\,[\ v\,(t_i)\ ] = 0, \quad E\,[\ v(t_i)\ \ v\,(t_j)] = 4\,\delta_{ij}, \quad (i, j = 0, 1, 2, \ldots) \tag{2.7}$$

where G_o is the dosage administered ar t_o.

Given Eqs. (2.2) - (2.7), the extended Kalman filter equations [4] are:

$$\dot{\hat{x}} = f\,[\ t,\ \hat{x}\,(t),\ D\,(t)\ ] \qquad t \in (t_{k-1}, t_k)\ (k=1, 2, \ldots) \tag{2.8}$$

$$\dot{S} = F\,[\ t, \hat{x}\,(t)\ ]\ \ S + S\,F^T\,[\ t,\ \hat{x}\,(t)\ ] \tag{2.9}$$

where

$$F\,[\ t, \hat{x}\,(t)\ ] = \partial f / \partial x\ \Big|_{x=\hat{x}} \tag{2.10}$$

and

$$K_k = S\,(t_k^-)\,H^T\,[\ \ H\,S\,(t_k^-)\,H^T + R_k\ ]^{-1} \tag{2.11}$$

$$S(t_k^+) = [\ \ I - K_k\,H\ ]\ \ S\,(t_k^-) \tag{2.12}$$

$$\hat{x}\,(t_k^+) = \hat{x}\,(t_k^-) + K_k\,[\ \ y\,(t_k) - x_3\,(t_k^-)\ ] \tag{2.13}$$

$$H = [\ \ 0\ \ \ 0\ \ \ 1\ \ \ 0\ \ \ 0\ \ \ 0\ \ \ 0\ ]\,. \tag{2.14}$$

In the equations above, t_k is a measurement time, R_k is the measurement noise covariance of Eq. (2.7), S is the state error covariance matrix, and

$$D(t) = \sum_{i=1}^{N}\ C(t_i)\ \ \delta\ (t - t_i). \tag{2.15}$$

The measurement times t_k $(k = 0, 1, \ldots)$ do not necessarily coincide with the dosage times t_i $(i = 0, 1, \ldots)$.

In Ref. 2 simulations of the above approach are presented. It is shown that approximately three measurements are required to either obtain a "reasonable" initial estimate of the parameter vector or detect a change in the patient's characteristics (which are implied by the parameter values). Also, an interactive graphics computer program containing the nonlinear filter has been developed for updating patient characteristics and determining optimal dosage strategies. The program is currently in a test phase at the University of Michigan Hospital.

3. OPTIMAL MEASUREMENT TIME FOR IDENTIFICATION

In the early stages of therapy, the prothrombin time measurements are useful for both evaluating the state of the patient and determining the patient-dependent parameters. In fact, the knowledge of the parameters is the main ingredient for the dosage optimization program to be described in Section 4. Thus, if possible, the early measurements (i.e., the Quick prothrombin times) should be made at times which are most beneficial for parameter identification. Some of the data employed in the study were obtained from the Thoracic Surgery Unit of the University of Michigan Hospital, and this Unit had the freedom to make the measurements at arbitrarily specified times.

For guidance in determining when to perform the measurement after the dosage, the following problem was analyzed. Given the system

$$\dot{x} = -K x \tag{3.1}$$

$$\dot{K} = 0 \tag{3.2}$$

$$y(t_1) = x(t_1) + v(t_1), \tag{3.3}$$

determine the time, t_1, to measure x so that the a posteriori variance of K is minimized. (This model is, of course, not the same as Eqs. (1.1) - (1.4), but it is viewed as a representation for the problem of identifying the largest time constant in the given system.) Given appropriate statistics and by employing the extended Kalman filter, this optimization problem can be solved analytically as follows.

Define $x_1 \equiv x$, $x_2 \equiv K$, and suppose the following quantities are given:

$$E[x(t_o)] = \hat{x}_o, E[(x(t_o) - \hat{x}(t_o))(x(t_o) - \hat{x}(t_o))^T] = \begin{bmatrix} S_{11}(t_o) & 0 \\ 0 & S_{22}(t_o) \end{bmatrix} \tag{3.4}$$

$$E[v(t_1)] = 0, \quad E[v(t_1)^2] = r. \tag{3.5}$$

Utilizing Eqs. (2.8) - (2.13), the extended Kalman filter equations for this system are

$$\dot{\hat{x}}_1 = -\hat{x}_1 \hat{x}_2, \qquad \hat{x}_1(t_o) = \hat{x}_{10} \qquad t \in [0, t_1) \tag{3.6}$$

$$\dot{\hat{x}}_2 = 0, \qquad \hat{x}_2(t_o) = \hat{x}_{20} \tag{3.7}$$

$$\dot{S}_{11} = -2\hat{x}_2 S_{11} - 2\hat{x}_1 S_{12} \tag{3.8}$$

$$\dot{S}_{12} = -\hat{x}_2 S_{12} - \hat{x}_1 S_{22} \qquad\qquad t \epsilon [\, 0, t_1 \,] \tag{3.9}$$

$$\dot{S}_{22} = 0 \tag{3.10}$$

$$S(t_1^+) = \begin{bmatrix} S_{11} - S_{11}^2/(r+S_{11}) & S_{12} - S_{11}S_{12}/(r+S_{11}) \\ S_{12} - S_{11}S_{12}/(r+S_{11}) & S_{22} - S_{12}^2/(r+S_{11}) \end{bmatrix}_{t_1^-} . \tag{3.11}$$

Equations (3.6) - (3.10) can be integrated analytically, and then $S_{22}(t_1^+)$, the quantity to be minimized, can be determined as a function of t_1:

$$S_{22}(t_1^+) = S_{22}(t_o)\left[1 - 1/\left(1 + \frac{r + S_{11}(t_o)\, e^{-2\hat{x}_{20} t_1}}{S_{22}(t_o)\, \hat{x}_{10}^2\, t_1^2\, e^{-2\hat{x}_{20} t_1}}\right)\right] \tag{3.12}$$

Thus, to minimize $S_{22}(t_1^+)$ is equivalent to minimizing

$$J(t_1) = \frac{r + S_{11}(t_o)\, e^{-2\hat{x}_{20} t_1}}{S_{22}(t_o)\, \hat{x}_{10}^2\, t_1^2\, e^{-2\hat{x}_{20} t_1}} . \tag{3.13}$$

Upon differentiation of J with respect to t_1 and simplification, the following equation defines the optimal value of t_1:

$$2\, r\, e^{2\hat{x}_{20} t_1^*} (t_1^* \hat{x}_{20} - 1) - 2\, S_{11}(t_o) = 0. \tag{3.14}$$

In the anticoagulant problem, the value of $S_{11}(t_o)$ (i.e., the variance of the initial state estimate) is usually assumed to be zero or very small, and thus,

$$t_1^* \approx 1/\hat{x}_{20}, \tag{3.15}$$

i.e., perform the measurement at the inverse of the a priori estimate of the time constant. The largest a priori time constant is K_g with a value on the order of six. Thus,

$$t_1^* \approx 4 \text{ hours} \tag{3.16}$$

is the time after the dosage recommended to perform the Quick prothrombin time test.

Integration of Eq. (3.9) implies

$$S_{12} = -S_{22}(t_o)\, \hat{x}_{10}\, t\, e^{-\hat{x}_{20} t} \tag{3.17}$$

and it is easily shown that the maximum value of $|S_{12}(t)|$ is achieved at $t = 1/\hat{x}_{20}$. Thus, if $S_{11}(t_o) = 0$, the optimal measurement time is the time of maximum correlation between x and K. From an intuitive point of view, note that the value of $t_1^* \epsilon (0, \infty)$ because a measurement at t = 0 gives no information about K, while a measurement at $t \to \infty$ is of little use since Eq. (3.1) is stable, i.e., $x \to 0$ independent of $K > 0$.

4. OPTIMAL LOADING DOSAGE

Given an estimate of the patient's state and parameters, then the major goal is to determine an optimal dosage strategy. This problem is usually divided into two parts: the loading phase (the time period when P is reduced from its normal value to the region of the desired value of P, i.e., P_f) and the maintenance phase ($P \approx P_f$ and very long time period).

In the loading phase it is desired to choose a dosage schedule which will cause the prothombin complex activity to decrease to a prespecified level as rapidly as possible without excessive under - or overshooting. After considering various characterizations of this problem, the following parameter optimization problem (where the dosages $C(t_o)$, $C(t_1), \ldots, C(t_N)$ are the unknown parameters) has been found to characterize the desired situation while allowing reliable, rapid optimization with a Davidon-Fletcher-Powell algorithm:

$$\text{MINIMIZE:} \quad J = t_\ell^2 + W_1 \int_{t_\ell}^{t_f} [P(t) - P_f]^2 \, dt \tag{4.1}$$

$$\text{SUBJECT TO: Eqs. (1.1) - (1.4)} \tag{4.3}$$

$$0 \leq C(t_i) \leq C_{Max} \qquad (i = 0, 1, \ldots, N) \tag{4.4}$$

where

$t_\ell \equiv$ time that P(t) first achieves P_{Max}, with $P_{Max} = P_f + \Delta P_f$ (the maximum allowable value of P(t) during maintenance).

$t_f \equiv$ prespecified time, sufficiently large enough to achieve the maintenance phase.

$W_1 \equiv$ relative weighting factor (the choice of this value is described below).

During maintenance it is desired to have

$$P_{MIN} \equiv P_f - \Delta P_f \leq P(t) \leq P_f + \Delta P_f \equiv P_{MAX}, \tag{4.5}$$

and originally this inequality was included in the optimization. However, by proper choice of W_1, the optimal solution can be guaranteed to satisfy this inequality, and the resultant optimization problem (without the inequality on P) can be solved more rapidly and reliably. The value, of W_1 employed in the computer program is

$$W_1 = \frac{18}{(\Delta P_f)^2\ (t_f-3).} \tag{4.6}$$

This value is determined as follows. During maintenance, P(t) can be roughly approximated by

$$P(t) \approx P_f + \Delta P_f \sin 2\pi\, t, \tag{4.7}$$

and thus

$$J \approx t_\ell^2 + W_1 \int_{t_\ell}^{t_f} [\ \Delta P_f \sin 2\pi\, t]^2\, dt$$

or,

$$J \approx t_\ell^2 + \frac{W_1\,(\Delta P_f)^2}{2}\ (t_f - t_\ell) + \frac{W_1(\Delta P_f)^2}{8\pi}\ [\ \sin 4\pi\, t_\ell - \sin 4\pi\, t_f] \tag{4.8}$$

Since $t_f - t_\ell \geq 3$ (at least, since t_f is chosen to insure that maintenance has been achieved) and ΔP_f is on the order of five, the last term is negligible, and thus

$$J \approx t_\ell^2 + \frac{W_1(\Delta P_f)^2}{2}\ (t_f - t_\ell). \tag{4.9}$$

Note that the regulation-type term, i. e., $\int_{t_\ell}^{t_f} [\ P(t) - P_f]^2\, dt$, is approximately linear in t_ℓ while the minimum time term is quadratic in t_ℓ. If W_1 is chosen to cause the regulation term to be of the same order of magnitude as t_ℓ^2 for a reasonable value of t_ℓ (say $t_\ell = 3$ days), then the use of t_ℓ^2 instead of t_ℓ for minimum time will not allow $t_\ell \rightarrow t_f$. Thus, equating the two terms in Eq. (4.9) with $t_\ell = 3$ and solving for W_1 gives the result in Eq. (4.6).

Typical results of the application of the optimization program are shown in Fig. 1. The initial iterate is always chosen to be a larger than normal dosage, e. g., 10 mg/day for eight days, to insure that $t_\ell \neq t_f$ on the first iterate. That is, the convergence is much more rapid and reliable when the algorithm has to reduce the dosage as opposed to having to increase to the optimum.

In Ref. 2, a first approach to dosage optimization in the maintenance phase is presented. This problem is much simpler in that a simple dosage strategy is desired (e. g., 2. 5 mg/day) since the dose is usually administered by the subject in a nonclinical setting. Of course, the loading optimization program can also be used for maintenance optimization by eliminating the t_ℓ^2 term in Eq. (4.1), but to date this has not been necessary.

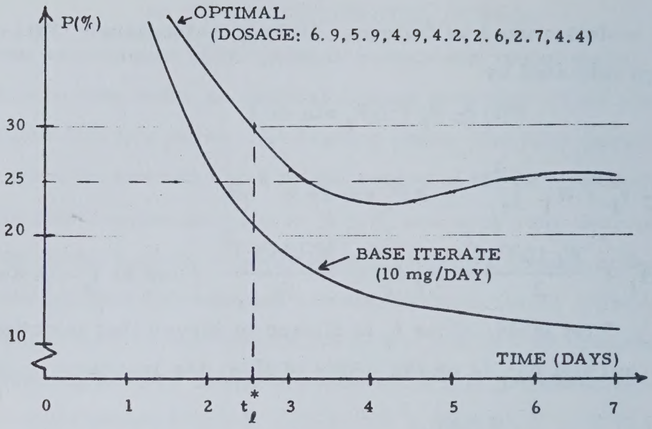


Figure 1. Optimal Seven-Day Loading Dosage.

5. CONCLUDING REMARKS

The loading phase of oral anticoagulant drug therapy has been attacked with modern state estimation and optimization techniques. The results to date have been encouraging, and an interactive graphics computer program employing the results is currently being tested in a clinical setting. Problems which remain to be studied involve a more thorough understanding of the maintenance phase and development of procedures for drug interactions.

REFERENCES

1. Abbrecht, P.H., and Huhta, J.C., "Clinical Applications of a Mathematical Model for Warfarin Dosage Determination", University of Michigan, Bioengineering Program Report, 1972.
2. Powers, W.F., Abbrecht, P.H., Covell, D.G., and Duleep, K.G., "Dynamic Systems Applications in Anticoagulant Therapy", AIAA Paper No. 76-804, presented at Nontraditional Applications Session of 1976 Astrodynamics Conference, San Diego, Calif., August 1976.
3. Quick, A.J., Hemorrhagic Diseases and Thrombosis, Lea & Febiger, Philadelphia, 1966.
4. Jazwinski, A.H., Stochastic Processes and Filtering Theory, Academic Press, New York, 1970.

RECURSIVE STATE OBSERVATION AND PARAMETER IDENTIFICATION USING HYPERSTABILITY THEORY

D.M. BOROSON AND S.C. SCHWARTZ
Dept. of Electrical Engineering and Computer Science
Princeton University, Princeton, New Jersey 08540

ABSTRACT

We consider the problem of simultaneously estimating the states and identifying the parameters of a time-invariant, discrete-time, linear system. Results from hyperstability are used in conjunction with a modified discrete adaptive observer. Conditions are established which guarantee that the recursive state estimation and parameter identification procedures converge in the noiseless case and an example is presented to illustrate the procedure.

I. INTRODUCTION

Although the separate problems of system parameter identification and state observation (estimation) have received considerable attention in the literature, the treatment of the joint problem has been less extensive. Simultaneously estimating states and system parameters, especially under noisy conditions, is much more difficult than estimating one set or the other, even if the system is linear, stable and time-invariant.

One recent approach has been to use the technique of model reference adaptive systems applied to input-output data [1]. Convergence of the recursive algorithms is demonstrated by generating a suitable Lyapunov function for the dynamical error equations. These procedures are well suited for continuous-time systems, but are more difficult to apply to discrete-time problems because of the complications involved in showing that a proposed Lyapunov function has negative definite forward differences. Kudva and Narendra have succeeded in constructing such a Lyapunov function and have applied it to their discrete adaptive observer ([2]). They have proved that, under noiseless conditions, the state and parameter estimates simultaneously converge to their true values when the input is chosen to excite all modes of the system. Their algorithm is simple, of fixed memory size, and recursive. One apparent limitation, however, seems to be that the recursion for parameter identification uses a fixed step size. Intuitively, for applications where noise disturbances are present, one would like the step size to decrease in some fashion, in a manner analogous to stochastic approximation type algorithms.

The notion of hyperstability introduced by Popov ([3]) has recently been used by Landau for the parameter identification problem ([4]). Here, a model reference adaptive system, in conjunction with results from hyperstability theory, lead to consistent parameter identification algorithms which have this feature of decreasing step size. Hyperstability is defined in the Appendix. Loosely speaking, the generalized energy in a hyperstable system is bounded by the initial energy in the system plus the energy delivered by the input. Also, as indicated in the Appendix, under suitable conditions, the definition of hyperstability reduces to that of passivity. For our present purposes, the importance of hyperstability theory is that it allows one to decouple a complicated system into somewhat simpler subsystems. Stability or convergence of the overall system can then be evaluated by studying the simpler components.

Landau's results suggest that his design procedure could be applied to a modified discrete adaptive observer of Kudva and Narendra. The goal would be to obtain an algorithm for the simultaneous estimation of system states and parameters which is simple and recursive, requires finite memory, and which can handle a wide class of inputs and some noise effect, i.e., have a decreasing step size. That is the subject of this paper.

In the next section we introduce notation and summarize the results of Kudva and Narendra. In Section 3, we discuss the application of hyperstability theory and proceed to decouple the overall system into two subsystems - a time-invariant feedforward system and a time-varying feedback system. Section 4 outlines the proof of stability and an example is presented illustrating the procedure. Comments and conclusions are given in Section 5.

II. MODEL REFERENCE ADAPTIVE OBSERVER ([2])

Consider a linear, time-invariant, discrete-time system

$$\begin{aligned} \underline{x}\,(k+1) &= A\underline{x}(k) + \underline{B}u(k) \\ y(k) &= \underline{C}^{*}\underline{x}(k) \end{aligned} \qquad (2.1)$$

where the scalar input and output (u(k) and y(k)) are given. The matrix A, and vectors B & C describe an unknown, completely observable and controllable n-th order system. It can be shown ([2]) that an equivalent representation is

$$\underline{x}\,(k+1) = \begin{bmatrix} \alpha & \underline{a}^{*} & \underline{b}^{*} \\ \underline{r} & F & 0 \\ 0 & 0 & F \end{bmatrix} \underline{x}(k) + \begin{bmatrix} \beta \\ 0 \\ \underline{r} \end{bmatrix} u(k) \qquad (2.2)$$

$$y(k) = [1,0,0,\ldots,0]\ \underline{x}(k)$$

The vectors $\underline{a},\underline{b}$, each with (n-1) components and the scalars α and β are unknown. The (n-1) square matrix F and the vector $\underline{r}$ must be a completely observable pair, but are otherwise free to be chosen. Once they are specified, $\underline{a},\underline{b},\alpha$, and β are determined.

Notice that we have gone from an n-th order system to one of order (2n-1). The reason one considers this non-minimal realization is that it leads to a convenient observer system. To this end, partition the state vector as

$$\underline{x}(k) = \begin{Bmatrix} x^{1}(k) \\ \underline{x}^{2}(k) \\ \underline{x}^{3}(k) \end{Bmatrix} \qquad (2.3)$$

Hence, $y(k) = x^{1}(k)$, and $\underline{x}^{2},\underline{x}^{3}$ are each (n-1) vectors. Now, all the unknown parameters in (2.2) enter only the one state x^{1}

which is observed directly. An adaptive observer system with similar structure is

$$\underline{\hat{x}}(k+1) = \begin{bmatrix} \hat{\alpha}(k) & \underline{\hat{a}}(k)^* & \underline{\hat{b}}(k)^* \\ \underline{r} & F & 0 \\ 0 & 0 & F \end{bmatrix} \underline{\hat{x}}(k) \tag{2.4}$$

$$+ \begin{bmatrix} \hat{\beta}(k) \\ 0 \\ \underline{r} \end{bmatrix} u(k) + \begin{bmatrix} \hat{\lambda}_1(k) \\ \underline{\hat{\lambda}}_2(k) \\ 0 \end{bmatrix} [\hat{x}^1(k) - y(k)]$$

where the circumflex denotes an estimate and we have set

$$\hat{\lambda}_1(k) = \lambda_1 - \hat{\alpha}(k) \tag{2.5}$$

$$\underline{\hat{\lambda}}_2(k) = \underline{\lambda}_2 - \underline{r} \tag{2.6}$$

Define the following quantities:

a) the unknown parameter vector

$$\underline{\varphi} = [\alpha, \underline{a}^*, \underline{b}^*, \beta]^* \tag{2.7}$$

b) a linear recursive parameter identification algorithm

$$\underline{\hat{\varphi}}(k+1) = \underline{\hat{\varphi}}(k) - P(k)\underline{f}(k) \tag{2.8}$$

where the vector $\underline{f}(k)$, which is a function of the data, and the square matrix $P(k)$ will be specified below.

c) the error in the parameter estimates

$$\underline{\tilde{\varphi}}(k) = \underline{\hat{\varphi}}(k) - \underline{\varphi} \tag{2.9}$$

d) the error in the state estimates

$$\underline{e}(k) = \underline{\hat{x}}(k) - \underline{x}(k) = [e^1(k), \underline{e}^2(k)^*, \underline{e}^3(k)^*]^* \tag{2.10}$$

e) the vector $\underline{q}$ with 2n components

$$\underline{q}(k) = \begin{pmatrix} y(k) \\ \underline{\hat{x}}^2(k) \\ \underline{\hat{x}}^3(k) \\ u(k) \end{pmatrix} \tag{2.11}$$

In [2], $\underline{\lambda}_2 = 0$, the vector $\underline{f}(k)$ is chosen to be

$$\underline{f}(k) = [e^1(k+1) - \lambda_1 e^1(k)]\, \underline{q}(k)/||\underline{q}(k)||^2 \tag{2.12}$$

and the matrix $P(k)$ is chosen to be the constant matrix P. Consequently, from (2.8), the parameter identification algorithm becomes

$$\underline{\hat{\varphi}}(k+1) = \underline{\hat{\varphi}}(k) - P\underline{q}(k) \frac{[e^1(k+1) - \lambda_1 e^1(k)]}{||\underline{q}(k)||^2} \tag{2.13}$$

This equation, in conjunction with the adaptive observer (2.4), constitutes the combined state estimation and parameter identification algorithms. Choosing the quantities $\underline{r}$, F, λ_1, and the matrix P, they guarantee by utilizing Lyapunov theory that the overall system is globally asymptotically stable; i.e., $\underline{\tilde{\varphi}}(k) \to 0$ and $\underline{e}(k) \to 0$. Their Lyapunov function is obtained by considering the dynamical error equations

$$\underline{e}(k+1) = \begin{bmatrix} \lambda_1 & \underline{a}^* & \underline{b}^* \\ \underline{\lambda}_2 & F & 0 \\ 0 & 0 & F \end{bmatrix} \underline{e}(k) + \begin{bmatrix} \tilde{\underline{\varphi}}(k)^* \underline{q}(k) \\ 0 \\ 0 \end{bmatrix} \qquad (2.14)$$

which follow from (2.1), (2.2), and (2.4)-(2.6). A quadratic function $\underline{v}^*(k)Q\underline{v}(k)$ is considered, where the vector is defined as

$$\underline{v}(k)^* = [e^1(k) - \underline{q}^*(k-1)\tilde{\varphi}(k-1), \underline{e}^2(k)^*, \underline{e}^3(k)^*, \tilde{\underline{\varphi}}(k-1)^*] \qquad (2.15)$$

The matrix Q is diagonal with elements $\delta, R, S, \gamma P^{-1}$, where δ, γ are scalars and R, S, P^{-1} are positive definite matrices. With suitable choices for these quantities, the quadratic form is negative semidefinite and, hence, the overall system is stable. They then show ([2], Appendix B) that if the input is "sufficiently general," one obtains the desired convergence: $\tilde{\varphi}(k) \to 0$, $\underline{e}(k) \to 0$. In their paper, the matrix P can be any Hermitian matrix satisfying $oI<P<2I$, where I is the identity matrix.

In summary then, one has to implement the adaptive observer system (2.4), and the parameter identification procedure (2.13). With appropriate selection of the various free parameters, the overall system can be shown to be asymptotically stable.

III. APPLICATION OF HYPERSTABILITY THEORY

The notion of hyperstability introduced by Popov ([3]), and expounded by Landau, appears to be especially useful for the joint estimation and identification problem. This is so because the theory allows one to decouple the joint problem into two subsystems. With certain stability attributes for each of these systems, we can then conclude that the overall system (the joint problem) is also stable. In addition, the hyperstability approach seems to be more systematic than Lyapunov theory since constructing Lyapunov functions is a difficult procedure for discrete systems with non-simple structure, despite the clever choice displayed in [2].

Our approach will be to combine the non-minimal realization structure for the adaptive observer in [2] with Landau's basic procedure for designing the entire system to be asymptotically stable. The main difference from the procedure outlined above will be the selection of the sequences P(k) and $\underline{f}(k)$.

In order to decouple the overall system, we define the following quantities:

a) N(z) is the z-transform of a filter whose input is $e^1(k)$ and output is w(k)

$$w(z) = N(z)e^1(z)$$

b) the transfer function for the first state in the error equation (2.14)

$$G(z) = [1 \;\; 0 \;\; 0] \left\{ zI - \begin{bmatrix} \lambda_1 & \underline{a}^* & \underline{b}^* \\ \underline{\lambda}_2 & F & 0 \\ 0 & 0 & F \end{bmatrix} \right\}^{-1} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad (3.2)$$

which reduces to

$$G(z) = \frac{1}{z - \lambda_1 - \underline{\lambda}_2^* [zI-F]^{-1}\underline{a}} \tag{3.3}$$

From the dynamical error equation, (2.14), it is clear that the quantity $w(k) = \underline{\tilde{\varphi}}(k)^* \underline{q}(k)$ can be viewed as the input to the error propagation system. Thus, from (3.1) and (3.3), we have the time-invariant system shown in Figure 1.

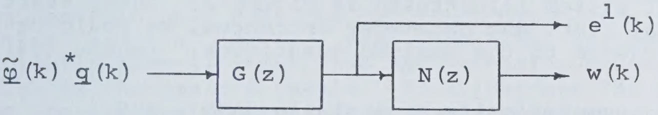


Figure 1

Similarly, from the parameter identification algorithm (2.8), $\underline{f}(k)$ can be viewed as the input to the linear parameter error propagation system

$$\underline{\tilde{\varphi}}(k+1) = \underline{\tilde{\varphi}}(k) - P(k)\,\underline{f}(k) \tag{3.4}$$

Hence, by choosing $\underline{f}(k)$ to be some function of $w(k)$ and $\underline{q}(k)$, we have defined the overall adaptive observer system which is readily decomposed into two subsystems:

1. a linear, time invariant system with input $\underline{\tilde{\varphi}}(k)^*\underline{q}(k)$ and output $w(k)$, which is in the forward path.

2. a linear, time-varying system with input $\underline{f}(k)$ and output $\underline{\tilde{\varphi}}(k)^*\underline{q}(k)$ which is in the feedback path.

The complete system is shown in Figure 2.[1]

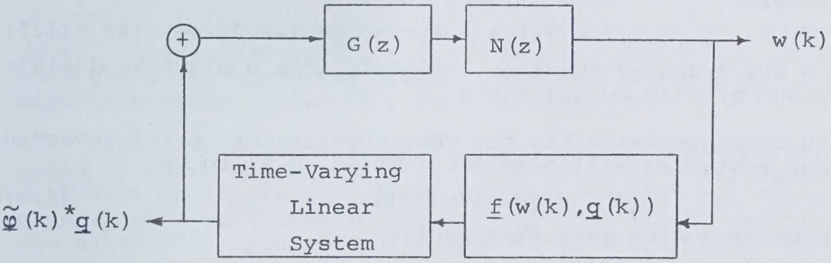

+
G(z)
N(z)
w(k)
Time-Varying Linear System
$\underline{f}(w(k),\underline{q}(k))$
$\underline{\tilde{\varphi}}(k)^*\underline{q}(k)$

Figure 2

We note that by choosing $N(z) = 1 - \lambda_1 z^{-1}$ and letting $\underline{f}(k) = \frac{\underline{q}(k)}{||\underline{q}(k)||^2}\, w(k+1)$ the previous algorithm is obtained (see 2.12). For $\underline{\lambda}_2 = 0$, we then have $G(z)N(z) = z^{-1}$. It was for these particular choices that the discrete Lyapunov function was obtained. In contrast, if we can now show that both $G(z)N(z)$ and the feedback system are hyperstable, then an important result of the theory is that the overall system is also hyperstable. Hence, we have a well-defined way of deciding how to filter the error to feed into the parameter equations, i.e., how to choose $N(z)$. Similarly, the gain matrices $P(k)$ can be chosen without concern for coupling effects.

---

[1] This representation is simplified in that we have omitted the effects of initial conditions.

IV. STABILITY ANALYSIS

Our goal is to show asymptotic stability in the large (ASIL) for the overall system illustrated in Figure 2. Then, starting with any initial state and parameter estimates, we would be assured of convergence to the desired quantities. Landau ([4]) has shown that if:

1. $G(z)N(z)$ is asymptotically hyperstable (AHS), and
2. the time-varying system is weakly hyperstable (WHS),

then the total system is ASIL.

The definitions of the different forms of hyperstability, as well as some related theorems, are given in the Appendix. Because of space limitations, only an outline of the stability analysis is given here. Details and proofs can be found in [5].

The time-invariant system is AHS if and only if it is strictly positive real (SPR), i.e., all poles of $G(z)N(z)$ are within the unit circle and $\mathrm{Re}[G(z)N(z)]>0$ for $|z| = 1$. (see [3]) The quantities $\lambda_1, \underline{\lambda}_2, N(z)$ are chosen to meet the SPR condition.[1] For example, we pick $\underline{\lambda}_2=0$ as in [2] and in Section 2 above, then $G(z) = 1/(z-\lambda_1)$. The simplest error filter which leads to a SPR transfer function is[2]

$$N(z) = (z-c)/(z-d) \tag{4.1}$$

For example,

$$(c,d,\lambda_1) = (-1.3,-.9,-.9) \text{ or } (-1.9,-.8,-.9) \tag{4.2}$$

yield a SPR transfer function. Clearly, one would like $G(z)N(z)$ to respond quickly to any input.

In order to establish the WHS property for the time-varying feedback system we will consider $f(k)$ in (3.4) to be

$$\underline{f}(k) = \underline{q}(k+1)w(k+1) \tag{4.3}$$

The feedback system must then satisfy

$$\sum_{j=0}^{K} w(j)\underline{q}(j)^{*}\underline{\tilde{\varphi}}(j) \geq -\gamma_0^2 \quad \forall K\geq 0 \tag{4.4}$$

where, in the definition of WHS, we are now considering $w(k)$ to be the input to the total feedback system. (See Figure 2 and (A.6).) One now substitutes for $\underline{\tilde{\varphi}} = \underline{\hat{\varphi}} - \underline{\varphi}$ using (3.4) and selects the gain matrix sequence $\{P(k)\}$ to ensure inequality (4.4). There are several ways to select such a sequence. (See [5])

---

[1] In addition $N(z)$ is restricted to be non-anticipatory so that the updating algorithm for $\underline{\tilde{\varphi}}(k)$ is realizable, i.e., $w(k)$ has to be a function of $e^1(k)$ and not $e^1(k+1)$. Thus, we require $G(z)N(z)$ to have strictly more poles than zeros when $\lambda_2=0$.

[2] By use of the bilinear transform, it is straightforward to show that the filter N=constant, and the one chosen in [2], $N(z)=1-\lambda_1 z^{-1}$, cannot lead to a SPR transfer function.

Having selected λ_1, λ_2, and $N(z)$ so that the time-invariant system is AHS, and having chosen the sequence $\{P(k)\}$ so that the time-varying system is WHS, we can now conclude that the overall system is hyperstable in **the** sense that the states of the time-invariant part satisfy Eq. (A.3). Using this fact, one can then show that the recursive procedures converge; i.e., $\underline{e}(k)\to 0$ and $\underline{q}(k)^*\underline{\tilde{\varphi}}(k)\to 0$. Then, under the assumption of a sufficiently general input, one can further show that $\underline{\tilde{\varphi}}(k)\to 0$, which is the desired result. To illustrate the procedure and the convergence properties, we now discuss an example.

The simulated system is of dimension two and is described by Equation (2.2). We have chosen

$$[\alpha, a, b, \beta, r, F] = [.6, -.5, 1.0, 1.0, 1.0, .5]$$

which results in a system transfer function

$$\frac{\beta z + (br-\beta F)}{z^2 - z(\alpha+F) + (\alpha F - ar)} = \frac{z + .5}{z^2 - 1.1z + .8} \qquad (4.5)$$

with a pair of complex poles at $.55 \pm j\sqrt{.4975}$. For our algorithm, we have selected the error filter

$$N(z) = (z + 1.9)/(z + .8) \qquad (4.6)$$

and $\lambda_1 = -.9$, $\lambda_2 = 0$. Two different $P(k)$ sequences were examined: $P(k)=P>0$, and (see [4])

$$P(k+1) = P(k) - \frac{P(k)q(k)q(k)^*P(k)}{1 + q(k)^*P(k)q(k)}, \quad P(0)>0 \qquad (4.7)$$

It is easy to show that the constant matrix ensures inequality (4.4). The second sequence is decreasing, which was our original goal. Although the proof in [4] is not complete, the algorithm seems to run very well in practice.

These two algorithms were compared to the algorithm proposed by Kudva and Narendra in [2]. The system was excited with various inputs including an iid uniformly distributed sequence and a sufficiently general periodic input. We also ran the algorithms using output measurements corrupted by another iid sequence. In every case, once the algorithms were "tuned", the new algorithm with $P(k)$ defined by (4.7) displayed a faster decrease in the state errors than the other two algorithms and an asymptotically better system parameter error. Similarly, using noise corrupted outputs, the new algorithm seemed to converge much faster (albeit to a biased result as in the algorithm given in [2]). By "tuned" we mean that, for all these algorithms the value of $P(0)$ greatly influenced both speed of convergence and initial overshoot. We found that, for $P(k)=P$, $.0001I \le P \le .001I$ exhibited the best behavior. For (4.7), $.01I \le P(0) \le .1I$ worked best and, for the algorithm of Kudva and Narendra, $.1I \le P \le I$ seemed to be optimal. When comparing different algorithms, we always made several runs with different choices of these initializing parameters and compared only the best results.

An advantage of algorithm (4.7) seemed to be that once it was tuned, and after an initial overshoot, the parameter errors decreased almost monotonically (under noiseless conditions) whereas the other two algorithms seemed to wander for a much longer time. In addition, a few other systems were tested (including an unstable one) and the results were similar. The simulations were run on an IBM 360/91 using Fortran IV and pseudo-random number generators. A typical result is presented in Figure 3.

V. CONCLUSION

Popov's development of hyperstability theory appears to have first been applied to the discrete system identification problem by Landau ([7]). In this paper, we have illustrated a similar design procedure for simultaneous state estimation and parameter identification. The technique incorporates a modified discrete adaptive observer and makes use of the results of hyperstability theory. Hyperstability properties allow one to decouple the overall system and analyze simpler subsystems. In this way, it would appear that a systematic approach can be developed for the design of the algorithms and their ultimate optimization in both deterministic and stochastic environments. (Recent results utilizing hyperstability in adaptive control have been reported in [8].)

Combining adaptive observers with hyperstability has a distinct advantage over Landau's design. In his algorithm, the choice of the error filter, N(z), depends on the system he is trying to identify. Specifically, he has to choose the numerator so that the transfer function is SPR when the denominator is the same as that of the unknown system he is trying to identify. Our algorithm with $\underline{\lambda}_2=0$ needs no knowledge of the actual system to guarantee the SP_R property. On the other hand, if we had chosen $\underline{\lambda}_2\neq 0$, from the definition of G(z), Eq. (3.3), we would be confronted with the similar problem of choosing N(z) so that G(z)* N(z) is SPR, while G(z) is a function of the parameters we are trying to identify.

It is apparent that a number of important areas still require investigation. Perhaps the most important are: simple (i.e., lower order) and convenient realizations of the adaptive observer, selection of the error filter N(z) and sequence P(k) to optimize convergence, analysis of the recursions in the presence of noise, optimal choice of inputs, and extensions to multi-input multi-output systems. At present, computer simulations make us optimistic for progress in many of these areas.

APPENDIX

In this Appendix we present the definitions of various forms of hyperstability and related results. A detailed account can be found in [3] and [9].

Given a linear discrete-time system

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) \\ y(k) &= C(k)x(k) + D(k)u(k) \ , \end{aligned} \qquad (A.1)$$

define the quadratic form

$$\eta(k_0,k_1) = \sum_{j=k_0}^{k_1} [u(j)^* x(j)^*] \begin{bmatrix} P(j) & Q(j) \\ Q(j)^* & R(j) \end{bmatrix} \begin{bmatrix} u(j) \\ x(j) \end{bmatrix} \qquad (A.2)$$

Def. 1: The system (A.1) (A.2) is said to be hyperstable (HS) iff $\exists\ \beta>0,\ \beta(k_0)\geq 0\ \ni$

$$\eta(k_0,k_1) + \beta(k_0)\|x(k_0)\|^2 \geq \beta\|x(k_1+1)\|^2, \forall k_1>k_0. \qquad (A.3)$$

Notice that for $P(j)=\frac{1}{2}(D(j)+D(j)^*), Q(j)=\frac{1}{2}C(j)$, η becomes

$$\eta(k_0,k_1) = \sum_{j=k_0}^{k_1} u(j)^* y(j) \qquad (A.4)$$

and for $\beta(k_0)=\beta(k)=1/2$, the inequality (A.3) reduces to

$$\eta(k_0,k_1) = \sum_{j=k_0}^{k_1} u(j)^* y(j) \geq \frac{1}{2}\|x(k_1+1)\|^2 - \frac{1}{2}\|x(k_o)\|^2. \qquad (A.5)$$

which is the definition of a passive system ([6]) with the energy interpretation mentioned in the introduction.

Def. 2: The system (A.1) is asymptotically hyperstable (AHS) iff it is HS and asymptotically stable in the large (ASIL).

Def. 3: Consider a system described by some input-output relation. This system is called weakly hyperstable (WHS) iff

$$\sum_{j=0}^{K} u(j)^* w(j) \geq -\gamma_0^2 \ \forall K\geq 0 \qquad (A.6)$$

where u(j) and w(j) are the input and output, respectively.

Let G_1, G_2 represent (possibly time-varying) linear systems. Let S_1 and S_2 represent the systems shown in Figs. 4 and 5, respectively

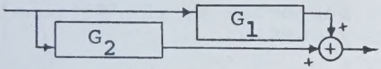


Figure 4

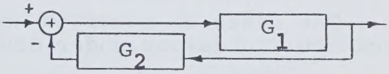


Figure 5

The following propositions can be found in [3]or[9], or derived from the above definitions.

Prop.1 HS implies WHS, assuming relation (A.4).
Prop.2 If G_1 and G_2 are WHS (HS) then system S_1 is WHS (HS).
Prop.3 If G_1 and G_2 are WHS (HS) then S_2 is WHS (HS).
Prop.4 If G_1 is HS and G_2 is WHS then S_2 is HS in the sense that the states of G_1 satisfy the inequality (A.5).
Prop.5 If G_1 is AHS and time-invariant and G_2 is WHS, then S_2 is ASIL in the sense that the states of $G_1 \to 0$ when the input=0.

The last two propositions are required in the stability analysis of Section 4.

An interpretation of hyperstability in terms of networks and its relationship to passivity can be found in [6]. See Popov [3] for extensions to more general systems.

REFERENCES

1. K.S. Narendra and L.S. Valavani, "Stable Adaptive Observers and Controllers", Proc. IEEE, Vol. 64, No.8, pp.1198-1208, Aug. 1976.

2. P. Kudva and K.S. Narendra, "An Identification Procedure for Discrete Multivariable Systems", IEEE Trans. on Auto. Contr., Vol. AC-19, No.5, pp.549-552, Oct. 1974. Also, see "The Discrete Adaptive Observer", Becton Center Tech. Report, CT-63, June 1974.

3. V.M. Popov, Hyperstability of Control Systems, New York: Springer-Verlag, 1973.

4. I.D. Landau, "Unbiased Recursive Identification Using Model Reference Adaptive Techniques", IEEE Trans. Auto. Contr., Vol. AC-21, No.2, pp.194-202, April 1976.

5. D.M. Boroson, "On System Identification and State Estimation in Linear, Discrete-Time Systems", Ph.D. Thesis in preparation, Princeton University, Princeton, N.J.

6. B.D.O. Anderson, "A Simplified Viewpoint of Hyperstability", IEEE Trans. Auto. Contr., Vol. AC-13, pp.292-294, June 1968.

7. I.D. Landau, "Synthesis of Discrete Model Reference Adaptive Systems", IEEE Trans. Auto. Contr., Vol. AC-16, pp.507-508, Oct. 1971.

8. J.M. Martin-Sanchez, "A New Solution to Adaptive Control", Proc. IEEE, Vol. 64, No.8, pp.1209-1218, Aug. 1976.

9. I.D. Landau, "Hyperstability Concepts and their Application to Discrete Control Systems", JACC 1972, pp.373-381.

ACKNOWLEDGEMENTS

The research reported here was supported by the National Science Foundation under Grant ENG-75-09610.

Legend for Figures:

3a: Solid Line - periodic input, RMS value = 2.088

Dotted Line - $\begin{cases}\text{periodic input, RMS value} = 2.088 \\ \text{output corrupted by iid noise} \sim U[-.5,.5]\end{cases}$

3b: Known stochastic input-iid, U[-2,2]

A: (2.13) , P= .1I

B: (4.3),(4.6) , P= .001I

C: (4.3),(4.6),(4.7) , $P(0) = \begin{cases}.01I & (3a) \\ .1I & (3b)\end{cases}$

$\hat{\underline{\phi}}(0)$ and $\hat{\underline{x}}(0)$ were $\underline{0}$ in all runs.

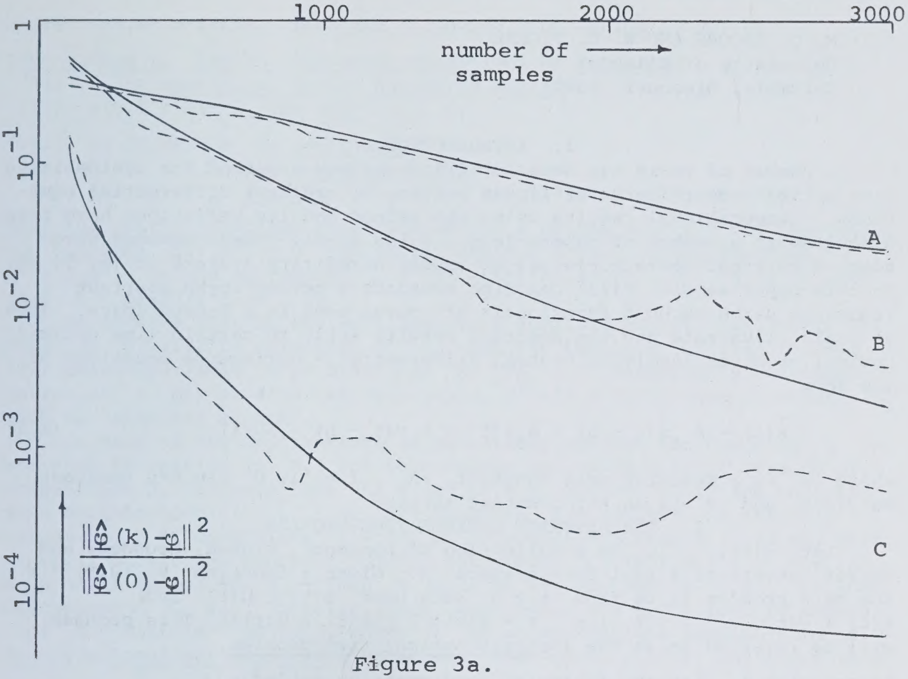


Figure 3a.

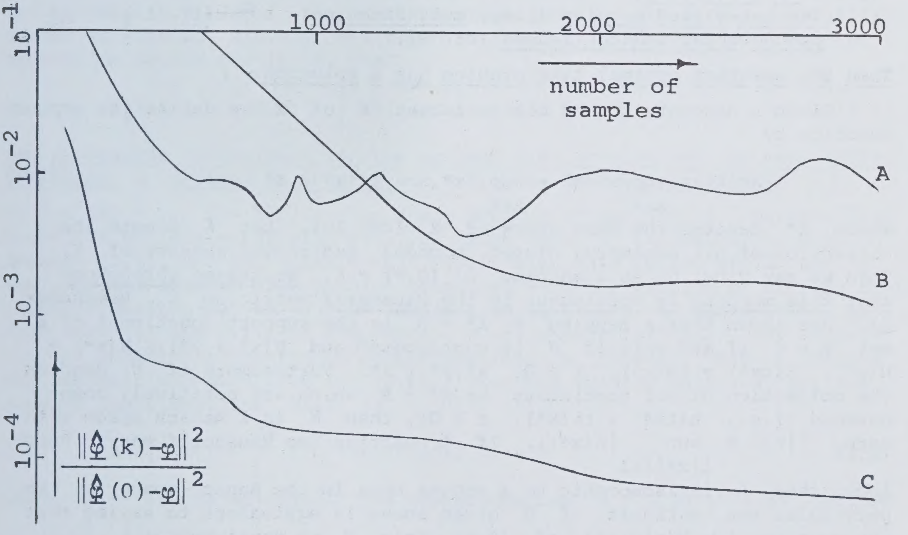


Figure 3b.

TIME OPTIMAL CONTROL OF HEREDITARY SYSTEMS

M. Q. JACOBS AND W. C. PICKEL
University of Missouri
Columbia, Missouri 65201

1. INTRODUCTION

A number of years ago Neustadt [1] described a method for synthesizing time optimal controllers for linear systems of ordinary differential equations. Computational results using the method and its variations have been discussed in a number of papers [e.g., 2, 3, 4, 5]. These methods were adapted to treat certain classes of linear hereditary systems in [6, 7, 8]. In this paper we will first describe Neustadt's method in an abstract framework using support functionals of convex sets in a Banach space. Then we shall illustrate how the abstract results apply to certain time optimal control problems involving neutral differential - difference equations of the form

$$\dot{x}(t) = A_{-1}\dot{x}(t-h) + A_0 x(t) + A_1 x(t-h) + Bu(t) \tag{1.1}$$

where h is a positive real constant, A_i , $i = \pm 1, 0$ are $n \times n$ constant matrices, and B is an $n \times r$ constant matrix.

2. THE ABSTRACT OPTIMAL TIME PROBLEM

Let $\Omega(t)$, $t \geq 0$ be a collection of nonempty, closed, bounded, and convex subsets of a real Banach space X. Given a function $z\colon [0,\infty) \to X$ the main problem is to find $\bar{t} \geq 0$ such that $z(\bar{t}) \in \Omega(\bar{t})$ and $z(t) \notin \Omega(t)$, $0 \leq t < \bar{t}$, i.e. $\bar{t} = \min\{t \geq 0 \mid z(t) \in \Omega(t)\}$. This problem will be referred to as the *abstract optimal time problem*.

Proposition 1. Let the following conditions be satisfied:

(i) $z(t_1) \in \Omega(t_1)$ *for some* $t_1 \geq 0$,

(ii) z *is continuous*,

(iii) *The set-valued mapping* (i.e., *multifunction*) $t \mapsto \Omega(t)$, $t \geq 0$ *is pseudo-upper semicontinuous* [cf. 9].

Then the abstract optimal time problem has a solution.

Given a nonempty closed convex subset K of X we define its support function by

$$H(x^*|K) = \sup_{x \in K} x^*(x) = \sup_{x \in K} \langle x^*, x\rangle \quad , \quad x^* \in X^* \tag{2.1}$$

where X^* denotes the dual space of X [cf. 10]. Let $\mathcal{K}$ denote the collection of all nonempty, closed, bounded, and convex subsets of X. Then we may view Ω as a mapping $\Omega\colon [0,\infty) \to \mathcal{K}$. *We assume throughout that this mapping is continuous in the Hausdorff metric on* $\mathcal{K}$. Hörmander [10] has shown that a mapping $H\colon X^* \to R$ is the support functional of a set $K \in \mathcal{K}$ if and only if H is continuous† and $H(x^* + y^*) \leq H(x^*) + H(y^*)$, $H(tx^*) = tH(x^*)$, $t \geq 0$, $x^*, y^* \in X^*$. Furthermore if $\mathcal{H}$ denotes the collection of all continuous $h\colon X^* \to R$ which are positively homogeneous (i.e., $h(tx^*) = th(x^*)$, $t \geq 0$), then $\mathcal{H}$ is a Banach space with norm, $\|h\| = \sup_{\|x^*\| \leq 1} |h(x^*)|$. If $\mathcal{K}$ carries the Hausdorff metric topology, then $\mathcal{K}$ is isomorphic to a convex cone in the Banach space $\mathcal{H}$. In particular the continuity of Ω given above is equivalent to saying that the mapping $t \mapsto H(\cdot|\Omega(t))$ of $[0,\infty)$ into $\mathcal{H}$ is continuous.

Define the function $g\colon [0,\infty) \times X^* \to R$ by

---

†The topology of X^* is the norm topology.

$$g(t,x^*) = H(x^*|\Omega(t)) - \langle x^*,z(t)\rangle \tag{2.2}$$

where $\langle x^*,x\rangle = x^*(x)$. Then the following proposition is valid.

Proposition 2. Let the following conditions be satisfied:

(i) Ω *is continuous in the Hausdorff metric;*

(ii) $z(t_1) \in \Omega(t_1)$ *for some* $t_1 \geq 0$;

(iii) $z\colon [0,\infty) \to X$ *is continuous;*

(iv) $\Omega(t)$, $t > 0$ *are strictly convex;*

(v) $0 \leq t_1 < t_2$ *implies* $g(t_1,x^*) < g(t_2,x^*)$, $x^* \in X^*$, $x^* \neq 0$.

Then the abstract optimal time problem has a unique solution $\bar{t}$. *Moreover, the equation* $g(t,x^*) = 0$ *implicitly defines a real valued function* $x^* \mapsto G(x^*)$ *on the half-space* $S = \{x^* \in X^* | \langle x^*,z(0)\rangle > 0\}$ *which is such that* $\bar{t} = \max\{G(x^*)|x^* \in S\}$.

We observe that in the more important applications of this result the function g will be of class C^1 so that the implicit function theorem [11] guarantees that $x^* \to G(x^*)$ is C^1 on S. Consequently the computation of $\bar{t}$ may be effected by solving $G'(x^*) = 0$ by some procedure such as "steepest ascent".

In many of the applications to hereditary systems the function g is not even increasing in t. For this reason one needs a local version of Proposition 2. Suppose that $H_0 = \{x^*|x^* \neq 0,\ g(\bar{t},x^*) = 0\}$ and that in some neighborhood of $\{\bar{t}\} \times H_0$ the function $t \mapsto g(t,x^*)$ is strictly increasing. If conditions (i)-(iv) of Proposition 2 remain in force, then we still obtain the function $G(x^*)$ satisfying $g(G(x^*),x^*) = 0$ for x^* in a neighborhood of H_0. Then $\bar{t}$ will be the maximum of $G(x^*)$ in some neighborhood of H_0. Moreover, these conditions are sufficient to assure that $\bar{t} = \max\{G(x^*)|x^* \in N\}$, where N is an appropriate neighborhood of H_0 is a *local optimal time* in the sense that $z(\bar{t}) \in \Omega(\bar{t})$ and there is a $\delta > 0$ such that for $\bar{t} - \delta \leq t < \bar{t}$, $z(t) \notin \Omega(t)$.

3. APPLICATIONS

Consider system (1.1) with initial function† $\phi \in W_2^{(1)}([-h,0],R^n)$ and $u \in L_2([0,t_1],R^r)$. The corresponding solution $x(t,\phi,u)$ of (1.1) can be written as $x(t,\phi,u) = x(t,\phi,0) + x(t,0,u)$. If $\Phi(t)$ denotes the transition matrix for (1.1), then

$$K_t(u) \equiv x(t,0,u) = \int_0^t \Phi(t-s)Bu(s)ds\ , \quad t \geq 0. \tag{3.1}$$

The admissible controllers for the optimal time problem are the measurable functions u *defined on intervals* $[0,t_1]$, $t_1 \geq 0$ *which satisfy* $-1 \leq u^i(t) \leq 1$, $i = 1,\ldots,r$ *on their domains of definition*. Define $\Omega(t)$ to be the collection of all $x(t,0,u)$ such that u is admissible with domain $[0,t]$, and let $z(t) = -x(t,\phi,0)$. With this choice of Ω and z the abstract optimal time problem coincides with the problem of driving the initial function ϕ to $0 \in R^n$ subject to (1.1) and u admissible. In this case (i) and (iii) of Proposition 2 are easily verified. For simplicity we restrict attention to scalar controls u $(r = 1)$. Let $\mathcal{U}$ denote the admissible controllers on $[0,t]$. Then one readily verifies that the support functional for $\Omega(t)$ is

$$H(\eta|\Omega(t)) = \int_0^t |K_t^*(\eta)(s)|ds, \quad t > 0 \tag{3.2}$$

where K_t^* is the adjoint of the transformation $K_t\colon L_2[0,t] \to R^n$ defined in (3.1), i.e.,

---

†For the sake of brevity we must refer the reader to [12] for any unexplained notation.

$$K_t^*(\eta)(s) = B^*\Phi^*(t - s)\eta\ , \quad \text{a.e. } s \in [0,t]. \tag{3.3}$$

It is not difficult to prove that $t \mapsto H(\eta|\Omega(t))$ is nondecreasing. To get that this function is strictly increasing in t is substantially more difficult and usually involves some sort of controllability hypothesis on (1.1). However, even when this function is strictly increasing the function $t \mapsto H(\eta|\Omega(t)) - \langle\eta, z(t)\rangle$ generally will not even be increasing. In this case one is frequently able to use the local results at the end of Section 2.

Another type of problem of considerable interest is the time optimal settling problem which has boundary conditions $x_0 = \phi$, $x_{t_1} = 0$ where t_1 is the terminal time. In this case $\Omega(t) = \{x_t(\cdot,0,u) | u \in \mathcal{U}\}$, $z(t) = -x_t(\cdot,\phi,0)$, $t \geq 0$. A suitable state space X of functions on $[-h,0]$ must be chosen. The support functional $H(x^*|\Omega(t))$ will be similar to (3.2) and this function will increase in the variable t. No general methods for synthesizing time optimal controllers for this problem are as yet available. However, in some cases (see Example 1) we are able to solve such problems by finite dimensional methods.

In the remainder of this paper we wish to discuss some solved examples when $n = 2$ and scalar controls $(r = 1)$ are used. The system matrices A_i, $i = \pm 1, 0$ and B are denoted by

$$A_i = \begin{bmatrix} \alpha_i\ , & \beta_i \\ \gamma_i\ , & \delta_i \end{bmatrix}\ , \quad i = \pm 1, 0\ , \quad B = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}\ . \tag{3.4}$$

We will use $[x,y]^T$ for the state vector in R^2. Necessary and sufficient conditions for (1.1) to be controllable in function state space $W_2^{(1)}([-h,0],R^n)$ were developed in [13]. These necessary conditions say that (1.1) is linearly equivalent to a system with coefficient matrices $\tilde{A}_i$, $i = \pm 1, 0$, $\tilde{B}$ where $\tilde{A}_{-1}$ is a companion matrix and $\tilde{B} = [0,1]^T$. We call such systems canonical systems. For simplicity of notation we assume that A_i, B, $i = \pm 1, 0$ are already in this linearly equivalent form. Then (1.1) is function space controllable if and only if either of the following two conditions is satisfied

$$\begin{aligned} &\text{(a)} \quad \beta_0 = 0 \quad \text{and} \quad \alpha_0 + \beta_1 + \alpha_1 e^{\beta_1 h} \neq 0 \\ &\text{(b)} \quad \beta_0 \neq 0 \quad \text{and} \quad \lambda_i + \beta_1 + \beta_0 e^{\lambda_i h} \neq 0\ , \quad i = 1,\ 2 \end{aligned} \tag{3.5}$$

where λ_i, $i = 1, 2$ are the roots of $\lambda^2 + (\beta_1 - \alpha_0)\lambda + \alpha_1\beta_0 - \alpha_0\beta_1 = 0$. For canonical controllable systems there is a feedback controller of the form

$$\begin{aligned} u(t) = v(t) &- (\gamma_{-1},\delta_{-1})(\dot{x}(t - h),\dot{y}(t - h))^T \\ &- (\alpha_1 + \gamma_0,\beta_1 + \delta_0)(x(t),y(t))^T - (\gamma_1,\delta_1)(x(t - h),y(t - h))^T \end{aligned} \tag{3.6}$$

which converts (1.1) to the system

$$\begin{aligned} \dot{x}(t) &= \dot{y}(t - h) + \alpha_0 x(t) + \beta_0 y(t) + \alpha_1 x(t - h) + \beta_1 y(t - h) \\ \dot{y}(t) &= -\alpha_1 x(t) - \beta_1 y(t) + v(t)\ . \end{aligned} \tag{3.7}$$

Time optimal control problems for neutral systems of the form (3.7) are relatively easy to solve by our methods. In fact such a system is equivalent to the following system

$$\begin{aligned} \dot{x}(t) &= \alpha_0 x(t) + \beta_0 y(t) + v(t - h) \\ \dot{y}(t) &= -\alpha_1 x(t) - \beta_1 y(t) + v(t) \ . \end{aligned} \tag{3.8}$$

In the time optimal control problem for (3.7) with boundary conditions $x_0 = \xi$, $y_0 = \eta$ $x(t_1) = y(t_1) = 0$, $\xi, \eta \in W_2^{(1)}[-h,0]$ the corresponding boundary conditions for (3.8) are $x(0) = \xi(0)$, $y(0) = \eta(0)$, and $v(t) = \dot{\eta}(t) + \alpha_1 \xi(t) + \beta_1 \eta(t)$, $-h \le t \le 0$, and $x(t_1) = y(t_1) = 0$. If we have the time optimal control problem for (3.7) with terminal condition $x_{t_1+h} = y_{t_1+h} = 0$, then we add to our conditions for (3.8) the requirement $v(t) = 0$, $t_1 - h \le t \le t_1$. In (3.7) or (3.8) the control restraint is $-1 \le v \le 1$ on $[0,t]$. Thus the problem (3.7) falls in the category of problems for systems of the form

$$\dot{w} = Aw + B_0 v(t) + B_1 v(t - h) \tag{3.9}$$

which were studied in great detail in [7, 8]. It is interesting to note that a fundamental aspect of the sufficiency theory [7, 8] for time optimal control of systems (3.9) is the condition $(n = 2)$ that the rank of $[B_0 + \exp(-Ah)B_1, AB_0 + A\exp(-Ah)B_1]$ be 2. For (3.7) this coincides exactly with the controllability conditions (3.5) for the neutral system (3.7). For systems in the form (3.9) we take

$$\begin{aligned} \Omega(t) &= \{\int_0^t \exp[-As][B_0 v(s) + B_1 v(s - h)]ds \mid v \ \underline{\text{admissible}}\}, \\ z(t) &= -[x(0), y(0)]^T, \quad t \ge 0, \end{aligned} \tag{3.10}$$

in the abstract optimal time problem. Effective methods for verifying the hypotheses of Proposition 2 are given in [7, 8]. Rather than describe these here we will conclude by giving examples of two time optimal problems of the type discussed here.

Example 1. The system equations are

$$\begin{aligned} \dot{x}(t) &= \dot{y}(t - 1) - x(t - 1) \\ \dot{y}(t) &= x(t) + u(t), \quad t \ge 0. \end{aligned}$$

The initial conditions are $x(t) = \xi$, $y(t) = \eta$, $-1 \le t \le 0$, where ξ, η are real constants. The terminal conditions are $x(t) = y(t) = 0$, $t_1 - 1 \le t \le t_1$. The controllers are required to satisfy $-1 \le u \le 1$. When this problem is converted to its equivalent form (3.9), the corre- Ω and z in (3.10) do satisfy the hypotheses of Proposition 2. Moreover if we let $\tau = T(\xi,\eta)$ denote the optimal time we get that the time optimal controller is

$$\left[\begin{array}{ll} -\mathrm{sgn}(\eta + \frac{\xi}{2}) & 0 \le t \le \frac{\tau}{2} - 1, \\ \mathrm{sgn}(\eta + \frac{\xi}{2}) & \frac{\tau}{2} - 1 < t \le \tau - 2, \\ 0 & \tau - 2 < t < \tau, \end{array}\right.$$

where $\tau = T(\xi,\eta) = 2 + \sqrt{|4\eta + 2\xi|}$.

Example 2. The system equations are

$$\begin{aligned} \dot{x}(t) &= -y(t) + v(t - h) \\ \dot{y}(t) &= x(t) + v(t), \end{aligned} \tag{3.11}$$

where $h = \frac{\pi}{3}$. This system also falls in the category of neutral problems mentioned in (3.7), (3.8). If we take $x(0) = \xi$, $y(0) = \eta$ and $x(t_1) = y(t_1) = 0$ as boundary conditions, and measurable admissible controllers satisfying $|v(t)| \le 1$, $-h \le t \le t_1$, then again we have an abstract optimal time problem with data as in (3.10). The sets $\Omega(t)$ are not even increasing for $t > h$. Nonetheless the local version of Proposition 2 does apply to this situation, and this problem was solved analytically in [8]. The solution is too complicated to describe here, but we can make a few observations about the solution by using the two figures below (Fig. 1, Fig. 2). In Fig. 1 we have plotted the sets $\Omega(t)$, $t = h = \frac{\pi}{3}, \frac{5\pi}{12}, \frac{\pi}{2}, 2\pi, \frac{25\pi}{12}$. Clearly these sets are not increasing. For $t < \frac{5\pi}{6}$ the boundaries of the reachable sets $\Omega(t)$ have corners, but for $t \ge \frac{5\pi}{6}$ these boundaries are smooth. In Fig. 2 we have plotted some typical optimal trajectories from the points $P_1(-.909,-.907)$, $P_2(1.637,-.458)$, and $P_3(1.862,-3.066)$ which are on the boundaries of $\Omega(\frac{\pi}{3})$, $\Omega(\frac{\pi}{2})$, $\Omega(2\pi)$ respectively. Each arc of an optimal trajectory is an arc of a circle with center $(\pm 1,\pm 1)$.

As mentioned above, this problem has been solved analytically in [8]. A numerical approach is also given there for problems involving a variety of boundary conditions on systems of the type (3.9). This procedure is described briefly below. Let

$$S(t) = \{x^* \mid g(t,x^*) < 0\}.$$

Choose $\varepsilon > 0$.

```
Step 1:  Set $i = 0$, $t_0 = 0$.
Step 2:  Set $i = i + 1$.
Step 3:  Choose $x_i^* \in S(t_{i-1})$.
Step 4:  Find the first time $t_i = G(x_i^*) > t_{i-1}$ such that $g(G(x_i^*),x_i^*) = 0$.
Step 5:  If $\mathrm{dist}(z(t_i),\Omega(t_i)) \ge \varepsilon$, go to step 2.
Step 6:  Set $\overline{x^*} = x_i^*$, $\overline{t} = G(\overline{x^*})$.
```

Although it may be nontrivial to check the condition in Step 5, the gradient of G, if it exists, may be used. This is the case locally in Example 2. If $||\nabla G(x_i^*)||$ is sufficiently small, then $\mathrm{dist}(z(t_i),\Omega(t_i)) \le \varepsilon$.

REFERENCES

1. L. W. Neustadt, "Synthesizing Time Optimal Control Systems", J. Math. Anal. Appl. 1(1960)[p. 484-493].

2. E. J. Fadden and E. G. Gilbert, "Computational Aspects of the Time Optimal Control Problem", Computing Methods in Optimization Problems, ed. A. V. Balakrishnan, L. W. Neustadt, Academic Press (1964), New York and London [p. 167-192].

3. E. G. Gilbert, "The Application of Hybrid Computers to the Iterative Solution of Optimal Control Problems", Computing Methods in Optimization Problems, ed. A. V. Balakrishnan, L. W. Neustadt, Academic Press (1964), New York and London [p. 261-284].

4. B. Paiewonsky, P. Woodrow, W. Brunner, and P. Halbert, "Synthesis of Optimal Controllers Using Hybrid Analog-Digital Computers", Computing Methods in Optimization Problems, ed. A. V. Balakrishnan, L. W. Neustadt, Academic Press (1964), New York and London [p. 285-303].

5. B. Paiewonsky, "Synthesis of Optimal Controls", Topics in Optimization, ed. G. Leitmann, Academic Press (1967), New York and London [p. 391-416].

6. M. N. Oguztoreli,"Time-Lag Control Systems",Academic Press (1966), New York and London [Ch. 10].

7. H. T. Banks, M. Q. Jacobs, and M. R. Latina, "The Synthesis of Optimal Controls for Linear Time Optimal Problems with Retarded Controls", JOTA, Vol. 8, No. 5(1971) [p. 319-366].

8. W. C. Pickel,"Numerical Methods for Time Optimal Control of Linear Systems with Applications to Neutral Systems",Ph.D. Thesis, University of Missouri-Columbia (1976).

9. M. Q. Jacobs, "Measurable Multivalued Mappings and Lusin's Theorem", Trans. Amer. Math. Soc., Vol. 134, No. 3(1968) [p. 471-481].

10. L. Hormander, "Sur la Function d'appui des Ensembles Convexes dans un Espace Localement Convexe", Arkiv for Matematik, Vol. 3, No. 12(1954) [p. 181-186].

11. D. G. Luenberger, "Optimization by Vector Space Methods", John Wiley, New York (1969).

12. H. T. Banks, M. Q. Jacobs, and C. E. Langenhop, "Characterization of the Controlled States in $W_2^{(1)}$ of Linear Hereditary Systems", Siam J. Control, Vol. 13, No. 3(1975) [p. 611-649].

13. M. Q. Jacobs and C. E. Langenhop, "Criteria For Function Space Controllability of Linear Neutral Systems", Siam J. Control and Opt. (to appear Nov. 1976).

This research was supported by the National Science Foundation under grant NSF MPS72 04695 A03.

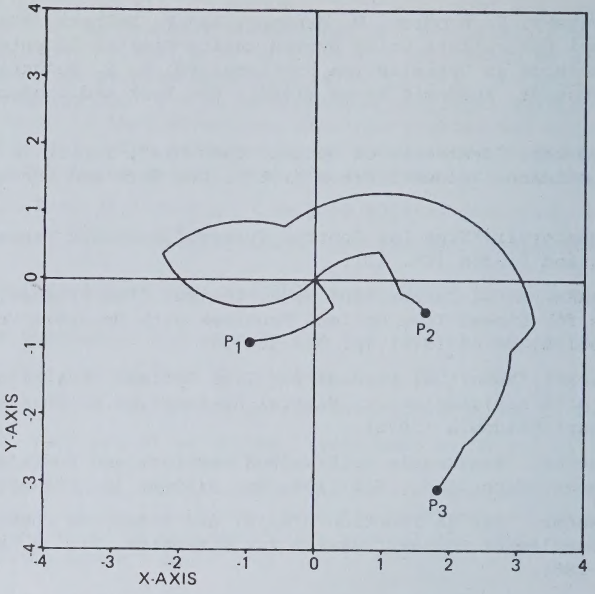

4
3
2
1
0
-1
-2
-3
-4
Y-AXIS
-4
-3
-2
-1
0
1
2
3
4
X-AXIS
P_1
P_2
P_3

Fig. 2

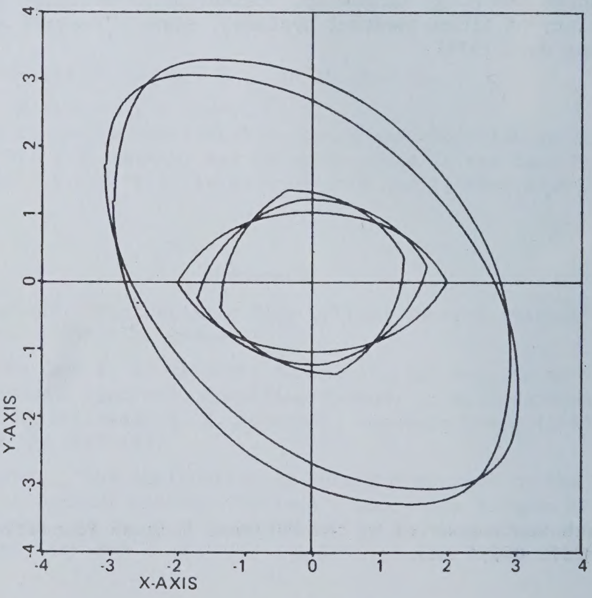

4
3
2
1
0
-1
-2
-3
-4
Y-AXIS
-4
-3
-2
-1
0
1
2
3
4
X-AXIS

Fig. 1

FILTERING OF LINEAR STOCHASTIC HEREDITARY DIFFERENTIAL SYSTEMS: NUMERICAL SOLUTION.*

MICHEL DELFOUR
Centre de Recherches Mathématiques, Université de Montréal
Montréal, P.Q., H3C 3J7, Canada.

ABSTRACT

Using an extention to hereditary differential systems of the well-known duality theorem of KALMAN-BUCY we give a numerical solution of the dual control problem. This method uses earlier results on the numerical solution of the operator Riccati differential equation (cf. M.C. DELFOUR [4]).

1. INTRODUCTION

The filtering problem for hereditary systems has been considered by a number of authors. To the author's knowledge, the first paper in this field is the one of H. KWAKERNAAK [1], where simultaneously the smoothing and filtering problems for linear differential systems with multiple constant time delays are studied. Other more recent papers by A. LINDQUIST [1], A. BENSOUSSAN [2], BENSOUSSAN-DELFOUR-MITTER [1], MITTER-VINTER [1], R. VINTER [2], R. CURTAIN [1], R. KWONG [1] and KWONG-WILLSKY [1] have also discussed the theory of this problem and extended the well-known duality theorem of KALMAN-BUCY in various forms.

In this paper we put our hereditary system in state form and use the work of A. BENSOUSSAN [1]. This leads to the study of the dual optimal control problem. It allows us to obtain the existence of the covariance operator $\Pi(t)$ and to study its properties without deriving the Riccati differential operator equation. One major difficulty is to make sense of that equation without adding any extra hypotheses on the matrices defining the original system. For instance R. VINTER [1] and MITTER-VINTER [1] have shown that the intersection over the time t of the domains of a certain unbounded operator $\tilde{A}(t)^*$ is generally not dense in the product space $X \times L^2(-a,0;X)$. Also the equation for the map r (cf. section 3) which appears in the decoupling of the optimality system (3.4)-(3.5) (cf. Theorem 3.1) does not belong to the class of hereditary systems; it could be interpreted as a special type of transport equation. But the above mentionned equations are not required to use J.C. NEDELEC [1]'s method (see also A. BENSOUSSAN [1]). As a result we obtain a numerical scheme to compute the covariance operator for which we have convergence proofs. A similar approach has been successfully used by M.C. DELFOUR [1] to [4] to numerically solve the Riccati differential equation describing the evolution of the feedback operator for the linear quadratic optimal control problem.

Several simple numerical examples are included to show the effect of delays on the operator $\Pi(t)$. More complete results can be found in M.C. DELFOUR [6]. They seem to generate discontinuities in the derivative of the map $\alpha \mapsto \Pi_{01}(t,\alpha)$. The covariance operator for this problem does not seem to have the same properties as the feedback operator in M.C. DELFOUR [1] to [4].

Notation. Given two real Hilbert spaces X and Y we denote by $\mathcal{L}(X,Y)$ the real Banach space of all continuous linear maps $L:X \to Y$ endowed with the natural norm $\|L\|$. The adjoint of L in $\mathcal{L}(X,Y)$ will be denoted by $L^* \in \mathcal{L}(Y,X)$.

---

* This research was supported in part by a FCAC Grant of the Quebec Ministry of Education and by National Research Council (Canada) Grant A8730 at the Centre de Recherches Mathématiques, Université de Montréal, Montréal, Canada.

When X=Y, we write $\mathcal{L}(X)$ and the identity in $\mathcal{L}(X)$ is denoted by I. An element L of $\mathcal{L}(X)$ is said to be self adjoint (resp. positive or ≥ 0) when $L^*=L$ (resp. for all x in X, the inner product of Lx and x in X is positive or zero).

Given a closed interval [a,b] and a real Banach space E, we denote by $L^p(a,b;E)$ the Banach space of all equivalence classes of m-measurable (m, the Lebesgue measure on [a,b]) maps $[a,b] \to E$ which are p-integrable ($1 \leq p < \infty$) or essentially bounded ($p=\infty$); its natural norm will be denoted $\| \ \|$. We shall also need the Sobolev space $H^1(a,b;E)$ of all maps x in $L^2(a,b;E)$ with a distributional derivative Dx in $L^2(a,b;E)$. C(a,b;E) will be the Banach space of all bounded continuous maps $[a,b] \to E$ endowed with the natural sup-norm. Finally $L^2_{loc}(0,\infty;E)$ will be the Fréchet space of all maps $[0,\infty[\to E$ the restriction of which to compact intervals of the form [0,T] belongs to $L^2(0,T;E)$.

2. SYSTEM DESCRIPTION AND FORMULATION OF THE PROBLEM.

2.1. Deterministic features.

Let X, U and Z be finite dimensional spaces, let N be a non-zero integer and let a and T be positive non-zero integers. Let $(\cdot,\cdot)$ and $|\cdot|$ denote the inner product and norm in X. The product space $H = X \times L(-a,0;X)$ is endowed with the following inner product and norm

$$(2.1) \qquad ((h,k)) = (h^0,k^0) + \int_{-a}^{0} (h^1(\theta),k^1(\theta))d\theta, \quad \|h\| = ((h,h))^{\frac{1}{2}}.$$

We are given real number $-a = \theta_N < \ldots < \theta_{i+1} < \theta_i < \ldots < \theta_0 = 0$ and maps

$$A_i:[0,\infty[\to \mathcal{L}(X), i=0,\ldots,N, \ B:[0,\infty[\to \mathcal{L}(U,X), \ A_{01}:[0,\infty[\times [-a,0] \to \mathcal{L}(X),$$

which are measurable and bounded on compact subsets. We now consider the hereditary system

$$(2.2) \quad \begin{cases} \dot{x}(t) = \sum_{i=0}^{N} A_i(t) \begin{Bmatrix} x(t+\theta_i) & ,t+\theta_i \geq 0 \\ h^1(t+\theta_i), & \text{otherwise} \end{Bmatrix} + \int_{-a}^{0} A_{01}(t,\theta) \begin{Bmatrix} x(t+\theta) & ,t+\theta \geq 0 \\ h^1(t+\theta), & \text{otherwise} \end{Bmatrix} d\theta \\ \qquad +B(t)\xi(t)+f(t) \text{ in }]0,T[\\ x(0)=h^0, \ h=(h^0,h^1) \in H, \ f \in L^2_{loc}(0,\infty;X), \ \xi \in L^2_{loc}(0,\infty;U). \end{cases}$$

We introduce the state $\tilde{x}(t)$ of system (2.2) at time t as follows

$$(2.3) \quad \tilde{x}(t) = (\tilde{x}(t)^0, \tilde{x}(t)^1) \in H, \ \tilde{x}^0(t) = x(t), \ \tilde{x}(t)^1(\theta) = \begin{Bmatrix} x(t+\theta) & ,-t \leq \theta \leq 0 \\ h^1(t+\theta), & -a \leq \theta < -t \end{Bmatrix}.$$

Let V be a subspace of H, $\Lambda: V \to H$ the canonical injection and $\tilde{A}(t):V \to H$

$$(2.4) \quad V=\{(h(0),h):h \in H^1(-a,0;X)\}, \tilde{A}(t)h = \left[\sum_{i=0}^{N} A_i(t)h(\theta_i) + \int_{-a}^{0} A_{01}(t,\theta)h(\theta)d\theta, \frac{dh}{d\theta} \right].$$

The state $\tilde{x}$ is the solution (in an appropriate space) of the state equation

$$(2.5) \quad \dot{\tilde{x}} = \tilde{A}\tilde{x} + \tilde{B}\xi + \tilde{f} \text{ in }]0,T[, \ \tilde{x}(0)=h, \ h \in V, \ f \in L^2_{loc}(0,\infty;X), \ \xi \in L^2_{loc}(0,\infty;U).$$

In a similar fashion, we can introduce the adjoint state equation

(2.6) $\dot{p} + \tilde{A}p + \Lambda^* g = 0$ in $]0,T[$, $p(T) = k \in H$, $g \in L^2(0,T;H)$.

2.2. Stochastic features.

We now consider a noisy initial condition, that is

(2.7) $x(0) = h^0 + \zeta^0$, $x(\theta) = h^1(\theta) + \zeta^1(\theta)$, $-a \leq \theta < 0$, $\zeta = (\zeta^0,\zeta^1) \in H$.

From now on ξ and ζ will be the noise at the input and the noise in the initial condition, respectively. We shall also assume an observation

(2.8) $$z(t) = C(t)x(t) + \eta(t),\ C:[0,T] \to \mathcal{L}(X,Z),$$

where C is measurable and bounded on compact intervals and η represents the error in measurement. As in A. BENSOUSSAN [1] $\{\zeta^0,\zeta^1,\xi,\eta\}$ will be modelled as a Gaussian linear random functional on the Hilbert space $\Phi = H \times L^2(0,T;U) \times L^2(0,T;Z)$ with zero mean and covariance operator

(2.9) $$\Xi = \begin{bmatrix} P_0 & 0 & 0 & 0 \\ 0 & P_1(\theta) & 0 & 0 \\ 0 & 0 & Q(t) & 0 \\ 0 & 0 & 0 & R(t) \end{bmatrix}.$$

It will be convenient to introduce the covariance operator P in $\mathcal{L}(H)$:

(2.10) $$((Ph,\bar{h})) = (P_0 h^0,\bar{h}^0) + \int_{-a}^{0} (P_1(\theta)h^1(\theta),\bar{h}^1(\theta))d\theta.$$

We can look at x(t) as a Gaussian linear random functional on X (for any t), where the mean of x(t), $\bar{x}(t)$, is a solution of (2.2) with $\xi=0$ and $\zeta=0$. It is easy to check that the mean of $\tilde{x}(t)$, $\bar{\tilde{x}}(t)$, is obtained from the respective pointwise means of x(t) and $h^1(\theta)$. As a result it is the solution of the state equation (2.5) with $\xi=0$ and $\zeta=0$ and the covariance operator of $\tilde{x}(t)$ is a solution of an integral analogue of a Lyapunov equation.

2.3. Formulation of the problem.

For each T we want to determine the best estimator of the linear random functional $\tilde{x}(T)$ with respect to the linear random functional z(s), $0 \leq s \leq T$. It is a linear random functional $\hat{\tilde{x}}(T)$ which can be obtained (see A. BENSOUSSAN [1]) through the following control problem. We start with the deterministic system (2.5) with the initial condition (2.7) at time 0 and consider the variables ξ and $\zeta = (\zeta^0,\zeta^1)$ as control variables. We want to minimize the cost function for a given h.

(2.11) $$J_T(\xi,\zeta,h) = ((P^{-1}\zeta,\zeta)) + \int_0^T [(Q^{-1}\xi,\xi)_U dt + (R^{-1}(z-Cx),z-Cx)_Z]dt$$

(provided that P, Q(t) and R(t) be invertible almost everywhere).

3. DUAL OPTIMAL CONTROL PROBLEM

We introduce the maps $g:[0,T] \to H$ and $N:[0,T] \to \mathcal{L}(H)$,

(3.1) $$g(t) = \tilde{C}(t)^* R(t)^{-1}[\tilde{C}(t)\bar{y}(t) - z(t)],\ N(t) = \tilde{B}(t)Q(t)\tilde{B}(t)^*,$$

where $\bar{y}$ is the solution of (2.5) with $\xi=0$ and $\tilde{x}(0) = h$. We consider the following control system and its associated cost function

(3.2) $\dot{p} + \tilde{A}^* p + \Lambda^*[\tilde{C}^* w + g] = 0$ in $]0,T[$, $p(T) = k \in H$, $w \in L^2(0,T;Z)$

$$(3.3) \qquad J^*(w,k) = (\!(Pp(0),p(0))\!) + \int_0^T [(\!(Np,p)\!)+(Rw,w)_Z]dt.$$

Given k, there exists a unique $\hat{w}$ which minimizes $J^*(w,k)$ over all w in $L^2(0,T;Z)$. It is characterized by the following optimality system:

$$(3.4)^* \qquad \dot{\hat{y}} = \tilde{A}\hat{y} - N\hat{p} \text{ in }]0,T[, \ \hat{y}(0) = -P\hat{p}(0), \ \hat{w} = R^{-1}\tilde{C}\hat{y},$$

$$(3.5) \qquad \dot{\hat{p}} + \tilde{A}^*\hat{p} + \Lambda^*[\tilde{C}^*R^{-1}\tilde{C}\hat{y}+g] = 0 \text{ in }]0,T[, \ \hat{p}(T) = k.$$

System (3.4)-(3.5) with k=0 is identical to the optimality system for problem (2.11) with the significant difference that (3.4)-(3.5) is derived without the hypothesis that P and Q(t) be invertible.

Remark. In the formulation of the dual optimal control problem $\tilde{C}$ can be any strongly measurable map $[0,T] \to \mathcal{L}(H,Z)$ which is bounded in [0,T]. It is not necessary to restrict ourselves to $\tilde{C}$'s of the form $\tilde{C}(t)h = C(t)h^0$ for some $C:[0,T] \to \mathcal{L}(X,Z)$.

We shall now proceed to the decoupling of optimality system (3.4)-(3.5) and to the study of the decoupling operator as in A. BENSOUSSAN [1].

In the sequel we shall use the notation

$$M(t) = \tilde{C}(t)^*R(t)^{-1}\tilde{C}(t).$$

Theorem 3.1. Let $\hat{p}$ and $\hat{y}$ be the solutions of system (3.4)-(3.5). Then there exists a family of linear operators $\Pi(t)$ in $\mathcal{L}(H)$ and a family of vectors r(t) in H, $0 \le t \le T$, such that

$$(3.6) \qquad \hat{y}(t) = -\Pi(t)\hat{p}(t) + r(t), \text{ in } [0,T].$$

The operator $\Pi(s)$ is a self adjoint positive element of $\mathcal{L}(H)$, and the map $s \mapsto \Pi(s):[0,T] \to \mathcal{L}(H)$ is continuous in the weak operator topology. It can be decomposed in a unique way into a matrix of operators

$$(3.7) \qquad \begin{bmatrix} \Pi_{00}(s), & \Pi_{01}(s) \\ \Pi_{10}(s), & \Pi_{11}(s) \end{bmatrix} \begin{matrix} \Pi_{00}(s) \in \mathcal{L}(X), \ \Pi_{01}(s) \in \mathcal{L}(L^2(-a,0;X),X) \\ \Pi_{10}(s) \in \mathcal{L}(X,L^2(-a,0;X)), \ \Pi_{11}(s) \in \mathcal{L}(L^2(-a,0;X)). \end{matrix} \quad \square$$

Formally one should obtain in]0,T[an equation of the form

$$(3.8) \qquad \dot{\Pi} = \Lambda^*\Pi\tilde{A} + \tilde{A}^*\Pi\Lambda - \Lambda^*\Pi M\Pi\Lambda + \Lambda^*N\Lambda, \ \Pi(0) = P.$$

However it is not easy to interpret this equation properly since $\tilde{A}(t)^*$ is an unbounded operator whose domain depends on time. In fact the intersection of the domains with respect to t is not necessarily dense in H (cf. R. VINTER [1] and MITTER-VINTER [1]). But (3.8) is meaningful in an integral sense.

4. APPROXIMATION OF THE DUAL SYSTEM.

In this section we shall exploit earlier results of M.C. DELFOUR [1] to [4] on the "primal" optimal control problem. We assume that we can find a discretization step $\delta > 0$ and non-zero positive integers

$$M,L,L_0,\dots,L_N \text{ such that } T = m\delta, \ a = L\delta, \ \theta_i = -L_i\delta \ (i=0,\dots,N).$$

4.1. Approximation of initial data.

We approximate the product space H by the finite dimensional space

---

* This equation must be interpreted in an appropriate weak sense.

(4.1) $$H^\delta = X^{L+1} \text{ with inner product } (\underline{h},\underline{k})_L = \sum_{\ell=-L}^{0} (h_\ell, k_\ell).$$

We introduce the maps

(4.2) $$h = (h^0,h^1) \mapsto r_\delta h = (h^0, h^1_{-1},\dots,h^1_{-L}) : H \to H^\delta,\quad h^1_\ell = \frac{1}{\delta}\int_{\ell\delta}^{(\ell+1)\delta} h^1(\theta)\,d\theta,$$

(4.3) $$\underline{h} = (h_0,h_{-1},\dots,h_{-L}) \mapsto q_\delta\underline{h} = \Big(h_0, \sum_{\ell=-L}^{-1} h_\ell\chi_\ell\Big) : H^\delta \to H,\quad I_\delta = q_\delta^* q_\delta : H^\delta \to H^\delta,$$

where χ_ℓ is the characteristic function of $[\ell\delta,(\ell+1)\delta[$.

4.2. Approximation of the differential equation.

We associate with A_i and A_{01} families of matrices for m=0,...,M-1

(4.4) $$A_i^m = \frac{1}{\delta}\int_{m\delta}^{(m+1)\delta} A_i(t)\,dt,\quad A_{01}^{m,\ell} = \frac{1}{\delta^2}\int_{m\delta}^{(m+1)\delta} dt \int_{\max\{(m+\ell)\delta-t,-a\}}^{\min\{(m+\ell+1)\delta-t,0\}} d\theta\, A_{01}(t,\theta),$$

for $\ell=0,\dots,-L$, and with f an element $\underline{f} = (f^0,f^1,\dots,f^{M-1})$ defined as

(4.5) $$f^m = \frac{1}{\delta}\int_{m\delta}^{(m+1)\delta} f(t)\,dt,\ m=0,\dots,M-1.$$

We construct a finite dimensional approximation $U^\delta = U^M$ to the space of control maps L (0,T;U). We define the map

(4.6) $$\xi \to r_\delta(\xi) = (\xi_0,\dots,\xi_{M-1}) : L^2(0,T;U) \to U^M,\quad \xi_m = \frac{1}{\delta}\int_{m\delta}^{(m+1)\delta} \xi(t)\,dt,$$

and a map q_δ similar to (4.2) and (4.3). It will be clear from the context whether q_δ and r_δ are associated with $L^2(0,T;U)$ or H. Finally we associate with B a family of matrices B^m in the same way as A_i^m in (4.4).

We define an $(L+1)n \times (L+1)n$ matrix $\tilde{A}^m$ and an $(L+1)n \times m$ matrix $\tilde{B}^m$

$$[\tilde{A}^m\underline{h}]_0 = \sum_{i=0}^{N} A_i^m h_{-L_i} + \delta \sum_{\ell=-L}^{0} A_{01}^{m,\ell} h_\ell,\quad [\tilde{A}^m\underline{h}]_\ell = \frac{1}{\delta}[h_{\ell+1}-h_\ell],\ \ell=-L,\dots,-1,$$

$$[\tilde{B}^m w]_0 = B^m w,\quad [\tilde{B}^m w]_\ell = 0,\ \ell=-L,\dots,-1.$$

We now consider the following numerical scheme for the state

(4.7) $$\begin{cases} \tilde{x}_{m+1} - \tilde{x}_m = \delta[\tilde{A}^m\tilde{x}_m + \tilde{B}^m\xi_m + \tilde{f}^m],\ m=0,\dots,M-1,\ \tilde{x}_0 = \underline{h}, \\ \underline{h} \in H^\delta,\ \xi = (\xi_0,\dots,\xi_{M-1}) \text{ and } \tilde{f} = (\tilde{f}^0,\dots,\tilde{f}^{M-1}),\ \tilde{f}^m = (f^m,0,\dots,0) \in H^\delta. \end{cases}$$

Proposition 4.1. Given h in H, ξ in $L^2(0,T;U)$ and f in $L^2(0,T;X)$ we denote by

$$(\tilde{x}_0,\dots,\tilde{x}_M) \in (H^\delta)^{M+1},\ \text{for } \underline{h} = r_\delta h,\ \underline{\xi} = r_\delta\xi,\ \underline{f} = r_\delta f,$$

the solution of (4.7) corresponding to $\underline{h}$, $\underline{\xi}$ and $\underline{f}$. We define the map $\tilde{x}^\delta$

(4.8) $$\tilde{x}^\delta(t) = \sum_{m=0}^{M-1} q_\delta\tilde{x}_m\chi_m(t),\ t \in [0,T[,\ \tilde{x}^\delta(T) = q_\delta\tilde{x}_M.$$

There exists a constant c > 0 (independent of δ, h, ξ and f such that

$$(4.9) \qquad \max\{\|q_\delta \tilde{x}_m\| : m=0,\dots,M\} + \|\tilde{x}^\delta\|_\infty \le c[\|h\|+\|f\|_2+\|\xi\|_2]$$

$$(4.10) \qquad \max\{\|q_\delta \tilde{x}_m - \tilde{x}(m\delta)\| : m=0,\dots,M\} \to 0 \text{ as } \delta \to 0 \ (M\delta=T),$$

where $\tilde{x}$ denotes the solution of equation (2.5). □

4.3. Approximation of the differential equation for the adjoint state.

In this section we introduce an approximation of the adjoint state equation (2.6). Consider the following scheme

$$(4.11) \quad p_{m+1} - p_m + \delta[I_\delta^{-1}(\tilde{A}^m)^* I_\delta p_{m+1} + g^m] = 0,\ m=0,\dots,M-1,\ p_M = \underline{k} \text{ in } H^\delta$$

where $\underline{k} = r_\delta(k)$ and $\underline{g} = (g^0,\dots,g^{M-1})$ is constructed from g in $L^2(0,T;H)$

$$(4.12) \qquad g^m = \frac{1}{\delta}\int_{m\delta}^{(m+1)\delta} r_\delta(g(t))\,dt,\ m=0,\dots,M-1.$$

Proposition 4.2. Given k in H, g in $L^2(0,T;H)$ and δ (small), we denote by $(p_0,\dots,p_M) \in (H^\delta)^{M+1}$, for $\underline{k} = r_\delta k$, $\underline{g} = (g^0,g^1,\dots,g^{M-1})$ as in (4.12), the solution of (4.11) corresponding to $\underline{k}$ and $\underline{g}$. We define the map p^δ

$$(4.13) \qquad p^\delta(t) = \sum_{m=0}^{M-1} q_\delta p_{m+1}\chi_m(t),\ 0 \le t < T,\ p^\delta(T) = q_\delta p_M.$$

There exists a constant c > 0 (independent of δ, k and g) such that

$$(4.14) \qquad \max\{\|q_\delta p_m\| : m=0,\dots,M\} + \|p^\delta\|_\infty \le c[\|k\|+\|g\|_2]$$

$$(4.15) \qquad \max\{|((q_\delta p_m - p(m\delta),h))| : m=0,\dots,M\} \to 0 \text{ as } \delta \to 0 \ (M\delta=T)$$

$$(4.16) \qquad \forall\ h, \{t \mapsto ((h,p^\delta(t)))\} \to \{t \mapsto ((h,p(t)))\} \text{ in } L^\infty(0,T;\mathbb{R})$$

where p is the solution of (2.6).

5. APPROXIMATION OF THE DUAL OPTIMAL CONTROL PROBLEM.

We now construct an approximation to the dual optimal control problem of section 3. We start with the approximation (4.11) to system (3.2)

$$(5.1) \quad \begin{cases} p_{m+1}-p_m+\delta[I_\delta^{-1}(\tilde{A}^m)^* I_\delta p_{m+1}+(\tilde{\underline{C}}^m)^* w_m+g^m] = 0,\ m=0,\dots,M-1, \text{ in } H^\delta \\ p_M=\underline{k}=r_\delta(k) \text{ for some } k \text{ in } H,\ \underline{w}=(w_0,\dots,w_{M-1})=r_\delta(w) \text{ for } w \in L^2(0,T;Z), \end{cases}$$

and associate with its solution the cost function

$$(5.2) \qquad J_\delta^*(\underline{w},\underline{k}) = (\underline{P}p_0,p_0)_L + \delta\sum_{m=0}^{M-1}[(\underline{N}^m p_{m+1},p_{m+1})_L + (\underline{R}^m w_m,w_m)_Z],$$

where g^m is defined as in (4.13), $\underline{R}^m$, $\underline{Q}^m$ as A_i^m in (4.4) and

(5.3) $\underline{P} = q_\delta^* P q_\delta,\ \underline{N}^m = \underline{\tilde{B}}^m \underline{Q}^m (\underline{\tilde{B}}^m)^*,\ \underline{\tilde{C}}^m = \frac{1}{\delta} \int_{m\delta}^{(m+1)\delta} \tilde{C}(t) q_\delta dt,\ m=0,\dots,M-1.$

The approximate optimal control problem consists in minimizing the cost function (5.2) over all $\underline{w}$ in Z^δ. It has a unique solution $\hat{w}$ (for δ small) which is completely characterized by the optimality system:

$$(5.4)\begin{cases} \hat{p}_{m+1}-\hat{p}_m+\delta[I_\delta^{-1}(\tilde{A}^m)^* I_\delta \hat{p}_{m+1}+(\tilde{C}^m)^*\hat{w}_m+g^m] = 0,\ 0 \le m \le M-1,\ \hat{p}_M = \underline{k}, \\ \hat{y}_{m+1}-\hat{y}_m = \delta[\tilde{A}^m \hat{y}_m - I_\delta^{-1}\underline{N}^m \hat{p}_{m+1}],\ \hat{y}_0 = I_\delta^{-1}\underline{P}\hat{p}_0,\ \hat{w}_m = (\underline{R}^m)^{-1}\underline{\tilde{C}}^m I_\delta \hat{y}_m,\ 0\le m\le M-1. \end{cases}$$

Proposition 5.1. (i) Given k in H and δ (small), the approximate optimal control problem (5.1)-(5.2) has a unique solution $\hat{\underline{w}}$ in U^δ. If $\hat{w}$ denotes the optimal control problem in the minimization problem (3.2)-(3.3), then as δ goes to zero

$$q_\delta(\hat{\underline{w}}) \to \hat{w} \text{ and } J_\delta(\hat{\underline{w}}, r_\delta h) \to J(\hat{w},h).$$

(ii) If $\hat{p}$ and $\hat{y}$ are the solutions of system (3.4)-(3.5), then

(5.5) $\forall\ s \in [0,T],\ q_\delta \hat{p}_m$ (resp. $\hat{y}_m$) $\to \hat{p}(s)$ (resp. $\hat{y}(s)$) in H weak with $w\delta=s$. □

Proposition 5.2. Let the sequences $\{\hat{p}_m\}$ and $\{\hat{y}_m\}$ be the solutions of the optimality equations (5.4). There exists a sequence of matrices $\{\Pi_m\}$ in $\mathcal{L}(H^\delta)$ and a sequence of vectors $\{\rho_m\}$ in H^δ such that

(5.6) $$I_\delta \hat{y}_m = -\Pi_m \hat{p}_m + \rho_m,\ m=0,\dots,M.$$

6. APPROXIMATION OF THE EQUATIONS FOR Π.

Theorem 6.1. As δ goes to zero

(6.1) $\exists\ c > 0$ (independent of δ), $\max\{\|r_\delta^* \Pi_m r_\delta\|_{\mathcal{L}(H)} : m=0,\dots,M\} \le c$

(6.2) $\forall\ h,k \in H,\ s \in [0,T],\ ((r_\delta^* \Pi_m r_\delta h,k)) \to ((\Pi(s)h,k))$ as $\delta \to 0$ $(m\delta=s)$.

Moreover, the sequence $\{\Pi_m\}$ is the solution of the following equations:

(6.3) $\Pi_{m+1} = \delta \underline{N}^m + (I+\delta I_\delta \tilde{A}^m I_\delta^{-1})\Pi_m (I+\delta \underline{M}^m \Pi_m)^{-1}(I+\delta I_\delta \tilde{A}^m I_\delta^{-1})^*,\ 0<m\le M,\ \Pi_0 = P.$ □

7. NUMERICAL EXAMPLES.

In this section we consider a number of examples which will illustrate the behaviour of the map $\alpha \mapsto \Pi_{01}(t,\alpha)$ as a function of the time t in [0,T]. In all examples a=1, T=2, $X=U=Z=\mathbb{R}$ and the observation equation is $z(t) = x(t) + \eta(t)$ with $R(t) = 1,\ 0 \le t \le 2$. All equations are to be interpreted in an appropriate way as stochastic differential equations.

Example 1. This example has an analytical solution. Consider

(7.1) $\dot{x}(t) = x(t-1),\ 0 \le t \le 2,\ x(0) = h^0 + \zeta^0,\ x(\theta) = h^1(\theta),\ -1 \le \theta < 0.$

Let $c = P_0^{-1}$. It can be shown that

$$(7.2)\qquad \Pi_{01}(s,\theta) = \left\{\begin{array}{ll} (s+c)^{-1}\chi_{[-s,0]}(\theta) & ,\ 0\le s\le 1 \\ s(1+c+\frac{s^3-1}{3})^{-1}\left\{\begin{array}{l} s+\theta,\ 1-s\le\theta\le 0 \\ 1\ ,\ \text{otherwise} \end{array}\right\},& 1\le s\le 2 \end{array}\right\}.$$

The results appear in Figures 1a and 1b for $P_0 = 1$.

Example 2. We now consider a system with two delays for $P_0 = 1$ (cf. Fig. 2)

(7.3) $\dot{x}(t) = x(t-\frac{1}{2})+x(t-1),\ 0 \leq t \leq 2,\ x(0) = h^0+\zeta^0,\ x(\theta) = h^1(\theta),\ -1 \leq \theta < 0.$

REFERENCES.

A. BENSOUSSAN [1], Filtrage optimal des systèmes linéaires, Dunod, Paris, 1971.

[2], Filtrage des systèmes linéaires avec retard, IRIA Report INF 7118/71027; oct. 1971.

A. BENSOUSSAN, M.C. DELFOUR and S.K. MITTER [1], Optimal filtering for linear stochastic hereditary differential systems, Proc. 1972 IEEE Conference on Decision and Control, New Orleans, 378-380.

R.T. CURTAIN [1], A Kalman-Bucy theory for affine hereditary differential equations, in Control theory, numerical methods and computer systems modelling, eds. A. Bensoussan and J.L. Lions, Springer-Verlag, New York 1975, 22-43.

M.C. DELFOUR [1], Solution numérique de l'équation différentielle de Riccati rencontrée en théorie de la commande optimale des systèmes héréditaires linéaires, in Control theory, numerical methods and computer systems modelling, eds. A. Bensoussan and J.L. Lions, Springer-Verlag, New York 1975, 362-383.

[2], Numerical solution of the optimal control problem for linear hereditary differential systems with a linear-quadratic cost function and approximation of the Riccati differential equation, CRM-Report 408, Université de Montréal, 1974.

[3], Numerical solution of the operational Riccati differential equation in the optimal control theory of linear hereditary differential systems with a linear-quadratic cost function, Proc. 1974 IEEE Conference on Decision and Control, Phoenix, 784-790.

[4], The linear quadratic optimal control problem for hereditary differential systems: theory and numerical solution, to appear in Journal of Applied Mathematics and Optimization.

[5], State theory of linear hereditary differential systems, to appear in J. Math. Anal. and Appl.

[6], Numerical solution of the operator Riccati equation for the filtering of linear stochastic hereditary differential systems, in Optimization techniques, modelling and optimization in the service of man, Part 2, ed. J. Céa, Springer-Verlag, New York 1976, 700-719.

H. KWAKERNAAK [1], Optimal filtering in linear systems with time delay, IEEE Trans. Automatic Control 12 (1967), 169-173.

R. KWONG [1], Doctoral dissertation, Massachusetts Institute of Technology, 1975.

[2], The linear quadratic Gaussian problem for systems with delays in the state, control, and observation, Proceeding of this conference.

R. KWONG and A.S. WILLSKY [1], Estimation and filter stability of stochastic delay systems, CRM-Report 613, Université de Montréal, 1976.

A. LINDQUIST [1], A theorem on duality between estimation and control for linear stochastic system with time delay, J. Math. Anal. Appl. 37 (1972), 516-536.

[2], Optimal control of linear stochastic systems with applications to time lag systems, Information Science 5 (1973), 81-126.

S.K. MITTER and R. VINTER [1], Filtering for linear stochastic hereditary differential systems, in Control theory, numerical methods and computer systems modelling, eds. A. Bensoussan and J.L. Lions, Springer-Verlag, New York 1975, 1-21.

J.C. NEDELEC [1], Schémas d'approximation pour des équations intégro-différentielles de Riccati, Thèse de doctorat d'état, Paris, 1970.

R. VINTER [1], On the evolution of the state of linear differential delay equations in M^2: properties of the generator, Report ESL-R-541, Electronic Systems Laboratory, M.I.T., Cambridge, Mass., U.S.A.
[2], A representation of solutions to stochastic delay equations, to appear.

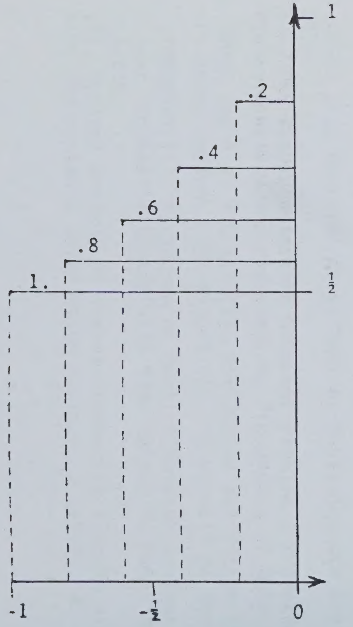


Figure 1a.

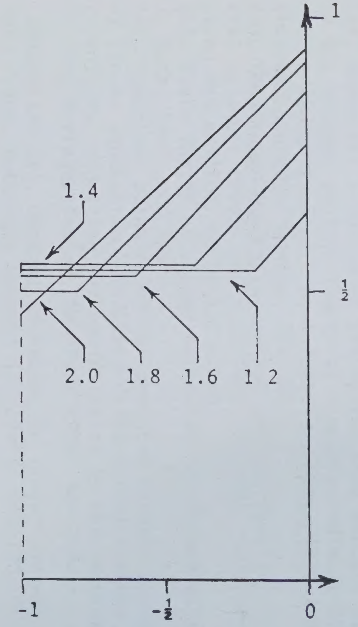
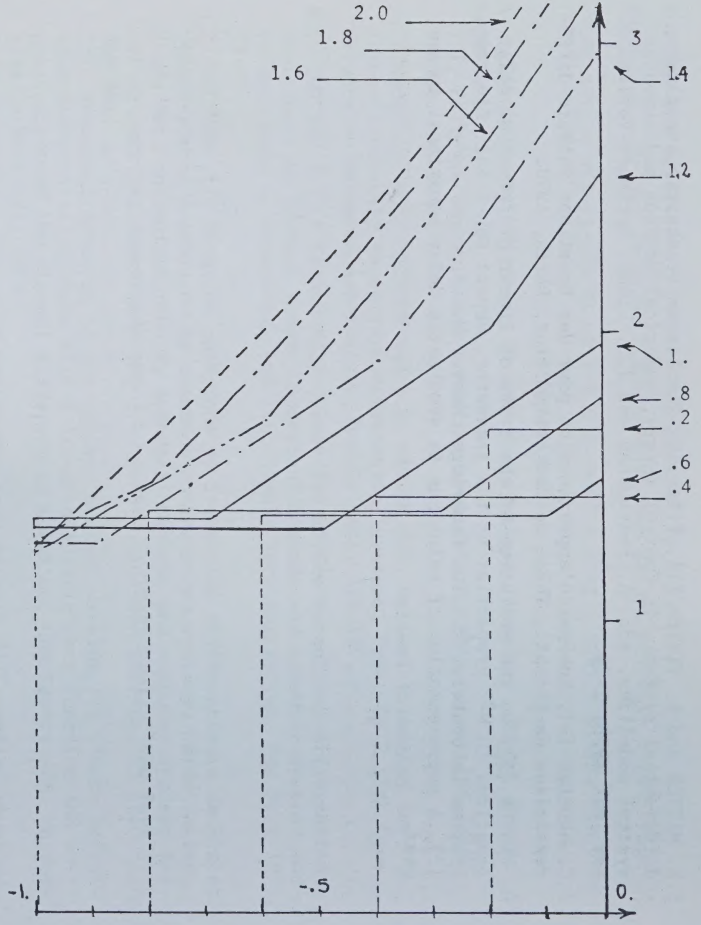


Figure 1b.

Figure 2.



THE LINEAR QUADRATIC GAUSSIAN PROBLEM FOR SYSTEMS WITH DELAYS IN THE STATE, CONTROL, AND OBSERVATIONS

R.H. KWONG
Department of Electrical Engineering, McGill University and Centre de Recherches Mathématiques, Université de Montréal, Montréal, Québec, Canada.

ABSTRACT

The linear quadratic Gaussian problem for systems with delays in the state, control, and observations is considered. The stability of the optimal time-invariant closed-loop system for the deterministic control problem is proved under the assumptions of stabilizability and observability. The stability of the optimal steady state filter for the stochastic estimation problem is then proved under the assumptions of detectability and controllability. Finally, these results are combined to prove the stability of the stochastic control system for the linear quadratic Gaussian problem.

INTRODUCTION

In previous papers [1] - [2], we have studied the linear quadratic Gaussian problem for systems with delays in the state only. A rather complete theory has been developed for this class of systems, including optimal control and filtering over finite and infinite time intervals, and the stability properties of the optimal systems. In this paper, we extend these results to linear stochastic systems with delays not only in the state, but also in the control and observations. The systems considered are of the form

$$dx(t) = [A_0 x(t) + A_1 x(t-h) + B_0 u(t) + B_1 u(t-h)]dt + F dw(t) \tag{1}$$

$$dz(t) = [C_0 x(t) + C_1 x(t-h)]dt + N dv(t) \tag{2}$$

with initial conditions $x(\theta) = 0$, $\theta \in [-h, 0]$, $u(\theta) = 0$, $\theta \in [-h, 0]$, and standard independent Brownian motion disturbances $w(t)$ and $v(t)$. The special initial conditions are not really essential restrictions but are imposed to simplify the already complex calculations. The cost functional is either of the form

$$J_T(u) = E \int_0^T [x'(t) H' H x(t) + u'(t) S u(t)] dt \tag{3}$$

or

$$J_\infty(u) = \lim_{T\to\infty} \frac{1}{T} E \int_0^T [x'(t) H' H x(t) + u'(t) S u(t)] dt \tag{4}$$

We shall obtain our results in several steps.

QUADRATIC OPTIMAL CONTROL

To study the above problem, we need to first solve the deterministic quadratic control problem and the stochastic filtering problem. The deterministic control problem is the following:

Given the linear delay system

$$\begin{aligned} \dot{x}(t) &= A_0 x(t) + A_1 x(t-h) + B_0 u(t) + B_1 u(t-h) \\ x(\theta) &= x_0(\theta) \qquad \theta \in [-h, 0] \end{aligned} \tag{5}$$

find the control law such that the cost functional

$$J_T^d(u) = \int_0^T [x'(t)H'Hx(t) + u'(t)Su(t)]dt \tag{6}$$

is minimized.

The case where $T < \infty$ has been studied by Koivo and Lee [3], who showed that the optimal feedback law is of the form

$$\begin{aligned} u(t) = -S^{-1}B_0'[L(t,t,t)x(t) &+ \int_{t-h}^{t} L(t,s+h,t)A_1x(s)ds \\ &+ \int_{t-h}^{t} L(t,s+h,t)B_1u(s)ds] - \gamma(t)S^{-1}B_1'[L(t+h,t,t)x(t) \\ &+ \int_{t-h}^{t} L(t+h,s+h,t)A_1x(s)ds + \int_{t-h}^{t} L(t+h,s+h,t)B_1u(s)ds] \end{aligned} \tag{7}$$

where $L(\sigma, s, t)$ is a resolvent kernel satisfying a certain integral equation, and $\gamma(t)=1$ if $0 \le t \le T-h$, $\gamma(t)=0$ if $T-h < t \le T$.

Our results here concern the infinite time problem where $T = \infty$. For the infinite time problem to be well-posed, we need the notion of stabilizability. The appropriate notion of stabilizability for systems of the form (5) is the following.

Definition 1. System (5) is said to be stabilizable if there exist matrix-valued functions K_0, $K_1(t)$ and $K_2(t)$ such that the feedback law

$$\begin{aligned} u(t) = K_0x(t) + \int_{-h}^{0} K_1(\theta)x(t+\theta)d\theta \\ + \int_{-h}^{t} K_2(t-s)x(s)ds \end{aligned} \tag{8}$$

stabilizes the system under consideration.

Compared to the stabilizability condition for systems with delays in the state only [4], the above definition contains an additional term involving the entire past state history $x(s)$, $s \in [0,t]$. The motivation for introducing the additional term in the present case is that we would like the control to be expressed solely in terms of feedback on the state. If we solve eq. (7) as an integral equation in u, we will obtain an expression for u of the form (8). This definition of stabilizability enables the dual notion of detectability for systems with delays in the observations to be formulated in a straight forward manner.

Stabilizability ensures that the infinite time problem is well-posed, but as usual, it does not guarantee that the optimal feedback law gives rise to an asymptotically stable system. We introduce here the notion of observability [1].

Definition 2. The system

$$\dot{x}(t) = A_0x(t) + A_1x(t-h) \tag{9}$$

$$z(t) = Hx(t) \tag{10}$$

is said to be observable if $z(t) = 0$ for all $t \ge 0$ implies $x(t) = 0$ for all $t \ge 0$.

Theorem 1. Assume that system (5) is stabilizable and system (9) - (10) is observable. Then the optimal control law for the infinite time quadratic control problem is given by

$$u(t) = -S^{-1} B_0' [L_{00} x(t) + \int_{-h}^{0} L_1(\theta) A_1 x(t+\theta) d\theta + \int_{-h}^{0} L_1(\theta) B_1 u(t+\theta) d\theta] - S^{-1} B_1' [L_{01} x(t) + \int_{-h}^{0} L_2(\theta) A_1 x(t+\theta) d\theta + \int_{-h}^{0} L_2(\theta) B_1 u(t+\theta) d\theta] \tag{11}$$

where L_{00}, L_{01}, $L_1(\theta)$ and $L_2(\theta)$ are obtained from the limiting forms of the function $L(\sigma, s, t)$ as $T \to \infty$. Furthermore, the closed-loop system is asymptotically stable.

OPTIMAL LINEAR FILTERING

We now consider the optimal linear filtering problem for stochastic systems with delays in the state and observations. That is, given the stochastic delay system

$$dx(t) = [A_0 x(t) + A_1 x(t-h)] dt + F dw(t) \tag{12}$$

$$x(\theta) = 0 \qquad \theta \in [-h, 0]$$

$$dz(t) = [C_0 x(t) + C_1 x(t-h)] dt + N dv(t) \tag{13}$$

characterize the least squares estimate of $x(t)$ based on the observations $z(s)$, $0 \leq s \leq t$. This situation is the dual to the linear quadratic control problem for systems with delays in the state and control studied above. When the time interval is finite, the solution of the filtering problem has been obtained in [1] and [2]. We showed there that the optimal estimate $\hat{x}(t/t)$ is characterized by the following set of equations

$$d\hat{x}(t|t) = [A_0 \hat{x}(t|t) + A_1 \hat{x}(t-h|t)]dt + [P(t,t,t)C_0' + P(t,t-h,t)C_1'] . R^{-1}[dz(t) - C_0 \hat{x}(t|t)dt - C_1 \hat{x}(t-h|t)dt] \tag{14}$$

$$\hat{x}(t-h|t) = \hat{x}(t-h|t-h) + \int_{t-h}^{t} [P(t-h,s,s) C_0' + P(t-h,s-h,s) C_1'] . R^{-1} [dz(s) - C_0 \hat{x}(s|s) ds - C_1 \hat{x}(s-h|s) ds] \tag{15}$$

where $P(\sigma, s, t)$ is the error covariance matrix $E\{[x(\sigma) - \hat{x}(\sigma|t)][x(s) - \hat{x}(s|t)]' | z^t\}$ and satisfies a certain integral equation "dual" to the integral equation for the control gain $L(\sigma, s, t)$. Here we are interested in the asymptotic behaviour of the optimal filter. For this, we introduce the notions of controllability and detectability.

Definition 3. The system

$$\dot{x}(t) = A_0 x(t) + A_1 x(t-h) + F u(t)$$
$$x(\theta) = 0 \qquad \theta \in [-h, 0] \tag{16}$$

is said to be (approximately) controllable if for any $\epsilon > 0$ and any pair $(\xi(0), A_1 \xi)$, where ξ is an L_2 function defined on $[-h, 0]$, there exist a time t and a control u_ϵ such that

$$\| x(t) - \xi(0) \| + \| A_1 x_t - A_1 \xi \|_{L_2} < \epsilon$$

(As usual x_t is the function defined on $[-h, 0]$ by $x_t(\theta) = x(t+\theta), \theta \in [-h, 0]$).

Definition 4. The system

$$\dot{x}(t) = A_o x(t) + A_1 x(t-h) \tag{17}$$

$$z(t) = C_o x(t) + C_1 x(t-h) \tag{18}$$

is said to be detectable if there exist matrix-valued functions K_0, $K_1(t)$ and $K_2(t)$ such that the system

$$\dot{x}(t) = A_o x(t) + A_1 x(t-h) + K_o z(t) + \int_{-h}^{o} K_1(\theta) z(t+\theta) d\theta + \int_{o}^{t} K_2(t-s) z(s) ds \tag{19}$$

is asymptotically stable.

Theorem 2. Assume that system (16) is controllable and system (17) - (18) detectable. Then the optimal steady state filter exists and is asymptotically stable.

STOCHASTIC CONTROL

Lindquist [5]- [6] has proved that a version of the Separation Theorem holds for systems of the form (1) and (2). By combining our results on control and filtering in the previous sections, we are now able to provide finally the remaining missing elements in the L-Q-G theory for general delay systems with delays in the state, control, and observations by characterizing the asymptotic behaviour of the closed-loop control system. Thus a complete L-Q-G theory for general linear delay systems is obtained.

Theorem 3. Assume that system (5) is stabilizable, system (9) - (10) is observable, system (16) is controllable, and system (17) - (18) is detectable. Then the control law given by

$$\begin{aligned} u(t) = & -S^{-1} B_o' [L_{oo} \hat{x}(t|t) + \int_{-h}^{o} L_1(\theta) A_1 \hat{x}(t+\theta|t) d\theta \\ & + \int_{-h}^{o} L_1(\theta) B_1 u(t+\theta) d\theta] - S^{-1} B_1' [L_{o1} \hat{x}(t|t) \\ & + \int_{-h}^{o} L_2(\theta) A_1 \hat{x}(t+\theta|t) d\theta + \int_{-h}^{o} L_2(\theta) B_1 u(t+\theta) d\theta] \end{aligned} \tag{19}$$

where $\hat{x}(t+\theta|t)$, $\theta \in [-h, 0]$, is the optimal steady state smoothed estimate of $x(t+\theta)$ given observations up to time t, generates an asymptotically stable system.

REFERENCES

[1] R.H. Kwong, "Structural properties and estimation of delay systems", Electronic Systems Lab. report ESL-R-614, Massachusetts Institute of Technology, Cambridge, Mass., Sept. 1975.

[2] R.H. Kwong and A.S. Willsky, "Estimation and filter stability of stochastic delay systems", submitted for publication.

[3] H. Koivo and E.B. Lee, "Controller synthesis for linear systems with retarded state and control variables and quadratic cost", Automatica, Vol. 8, 1972, pp. 203-208.

[4] L. Pandolfi, "On feedback stabilization of functional differential equations," Bolletino U.M.I., to appear.

[5] A. Lindquist, "Optimal control of linear stochastic systems with applications to time lag systems, "Information Science, Vol. 5, 1973, pp. 81-126.

[6] A. Lindquist, "On feedback control of linear stochastic systems," SIAM J Control, Vol. 11, May 1973, pp. 323-343.

REGULATOR THEORY FOR EVOLUTION SYSTEMS

K.P.M. Bhat
Dept. of Electrical Engineering
University of Toronto
Toronto M5S 1A4
Canada

ABSTRACT

Some aspects of the regulator theory of finite dimensional linear multivariable systems are studied for evolution systems defined in Banach spaces. Conditions are obtained for the existence of a regulating controller. A limited version of the Internal Model Principle is verified to hold for evolution systems. The general results are applied to time delay systems to obtain simple conditions in terms of the system parameters.

1. INTRODUCTION

In this paper we present some aspects of regulator theory for evolution systems. In particular a limited version of the Internal Model Principle (IMP) will be established for a very general class of linear systems which includes distributed parameter systems, hereditary systems, and, of course, finite dimensional systems. Recall here that the IMP asserts certain necessary structure of a controller which achieves regulation of the given system (for a precise statement see [1]). It was conjectured in [1] and [2] that the IMP is valid in more general situations than the finite dimensional case. Here we present partial results towards the verification of this statement. Other major works on regulator theory for finite dimensional systems are [3], [4], [5].

1.1 Notation

Vector spaces and their subspaces will be denoted by script letters $\mathcal{X}$, $\mathcal{Y}$ etc. and linear operators by roman capitals A, B, ... etc. For an operator $(\cdot)$, $\mathrm{Im}(\cdot)$ and $\mathrm{Ker}(\cdot)$ are respectively its range and null spaces. $\sigma(\cdot)$ is the spectrum of $(\cdot)$. $\mathbb{R}$ is the field of real numbers and $\mathbb{C}$ the complex plane. 0 stands for anything that is zero (a number, element of a space, a set or an operator), according to context.

1.2 System Description

The system we consider here is described by the following linear, time invariant evolution equation

$$\dot{x}_1(t) = A_1x_1(t) + B_1u(t) + A_3x_2(t), \quad t\geq 0 \qquad x_1(0) \in \mathcal{D}(A_1) \tag{1a}$$

Here A_1 is a closed operator defined on a dense subset $\mathcal{D}(A_1)$ of the Banach space $\mathcal{X}_1$ and generates a strongly continuous semigroup of bounded operators $\{T_{A_1}(t),\ t\geq 0\}$. The input $u(t)\in\mathcal{U}$, a Banach space and $x_2(t)$, the exosystem variable, satisfies

$$\dot{x}_2(t) = A_2x_2(t), \quad x_2(0)\in\mathcal{X}_2 \tag{1b}$$

and represents reference and/or disturbance signals which the regulator is to be designed to process. X_2 is a Banach space. The operators B_1, A_3 and A_2 are then appropriately defined and are assumed to be bounded. Without loss of generality we also assume $\sigma(A_2) \subset \mathbb{C}^+$.

The variables to be regulated are defined as

$$y(t) = C_1 x_1(t) + C_2 x_2(t) \tag{1c}$$

where $y(t) \in Y$, a Banach space and C_1 and C_2 are bounded operators.

Since A_2 is bounded, $x_2(t)$, the solution of (1b) is continuously differentiable and hence for each $x_1(0) \in \mathcal{D}(A_1)$ and for continuously differentiable input u(t), equation (1a) has a unique (strong) solution [6, 7]. Weak or generalized solutions of (1) may also be defined by allowing more general class of inputs [8].

2. REGULATION PROBLEM

The regulation problem is concerned with the synthesis of a controller in order to achieve internal stability and output regulation. The controller itself is an evolution system processing the output y:

$$\begin{aligned} \dot{x}_c(t) &= A_c x_c(t) + B_c y(t), \quad t \geq 0, \\ & x_c(0) \in \mathcal{D}(A_c), \end{aligned} \tag{2a}$$

$$u(t) = F_c x_c(t) + G_c y(t) \tag{2b}$$

The space X_c is again a Banach space and A_c is an infinitesimal generator of a strongly continuous semigroup. B_c, F_c and G_c are bounded operators between appropriate spaces. Thus, the controller is characterized by the 5-tuple $(X_c, A_c, B_c, F_c, G_c)$.

The closed loop system is now defined in the product space $X_L = X_1 \times X_c$ as

$$\begin{aligned} \dot{x}_L(t) &= A_L x_L(t) + B_L x_L(t), \quad t \geq 0, \\ x_L(0) &= (x_1(0), x_c(0) \in \mathcal{D}(A_1) \times \mathcal{D}(A_c), \end{aligned} \tag{3a}$$

$$y(t) = C_L x_L(t) + C_2 x_2(t) \tag{3b}$$

where

$$x_L(t) = \begin{bmatrix} x_1(t) \\ x_c(t) \end{bmatrix},$$

$$A_L = \begin{bmatrix} A_1 + B_1 G_c C_1 & B_1 F_c \\ B_c C_1 & A_c \end{bmatrix} : \mathcal{D}(A_L) \to X_L,$$

$$\mathcal{D}(A_L) = \mathcal{D}(A_1) \times \mathcal{D}(A_c)$$

$$B_L = \begin{bmatrix} A_3+B_1G_cC_2 \\ B_cC_2 \end{bmatrix} : X_2 \to X_L$$

and

$$C_L = [C_1 \quad 0] : X_L \to Y$$

A schematic flow graph for the closed loop system is in Figure 1.

Remark: Observe that in the above formulation the controller has the regulated variables y(t) as its input. Thus a feedback structure is assumed to be already present.

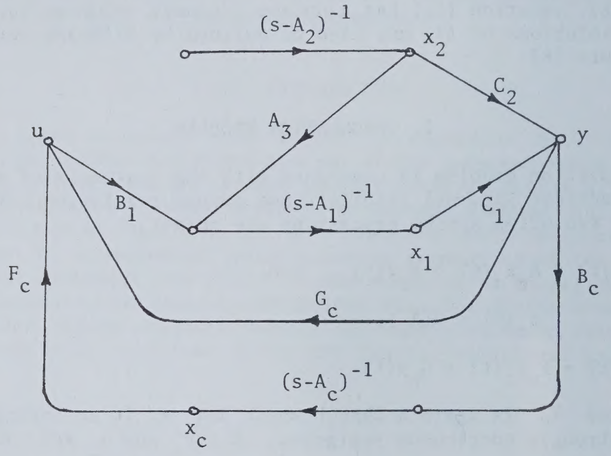


Fig. 1

Before we proceed to define the regulation problem in precise terms we need the following definition.

2.1 Definition

The infinitesimal generator A_1 (or the semigroup $T_{A_1}(t)$) is said to be stable if there exist positive constants M and γ such that

$$\| T_A(t) \| \le M e^{-\gamma t}, \quad t \ge 0.$$

The regulator problem consists of establishing conditions for the existence of a controller the purpose of which is to achieve

i) Internal stability: The infinitesimal generator A_L of the closed loop system is stable, and

ii) Output regulation: The output $y(t) \to 0$ as $t \to \infty$ for all $x_L(0) \in X_L$ and for all $x_2(0) \in X_2$.

The 5-tuple $(X_c, A_c, B_c, F_c, G_c)$ solving this problem will be called a synthesis and will be denoted for short by Σ_c.

3. TECHNICAL PRELIMINARIES

First, we need the following definition

3.1 Definition

The pair (A_1, B_1) is stabilizable if there exists a bounded operator $F_1 : X_1 \to \mathcal{U}$ such that $(A_1+B_1F_1)$ is stable. Similarly, the pair (C_1, A_1) is detectable if there exists a bounded operator $K_1 : \mathcal{Y} \to X_1$, such that $(A_1+K_1C_1)$ is stable.

We then obtain

3.2 Proposition

The stabilizability of (A_1, B_1) and detectability of (C_1, A_1) are both necessary for the stability of the closed loop system.

In view of this, we shall assume henceforth that (A_1, B_1) and (C_1, A_1) are respectively stabilizable and detectable.

In obtaining the solution of the regulation problem the following notation will be used. Let A be a densely defined closed operator in the Banach space $\mathcal{X}$ with domain $\mathcal{D}(A)$. Denote by $\underline{X}$ the set of all bounded linear operators from $\mathcal{X}_2$ into $\mathcal{X}$. Now define $\underline{A} : \mathcal{D}(\underline{A}) \to \underline{X}$ as $\underline{A}X := AX - XA_2$, $X \in \mathcal{D}(\underline{A})$. $\mathcal{D}(\underline{A})$ consists of all bounded linear operators from $\mathcal{X}_2$ into $\mathcal{D}(A)$. Finally, for any bounded linear operator $R : \mathcal{X} \to \mathcal{W}$, where $\mathcal{W}$ is a Banach space different from $\mathcal{X}$, define $\underline{R} : \underline{X} \to \mathcal{W}$ by $\underline{R}X = RX$, $X \in \underline{X}$. With this notation, we state the following result from [7] where it is established for the case when A and A_2 are both unbounded.

3.3 Lemma [7, p. 316]

If the spectra of A and A_2 in the extended complex plane* do not intersect, the equation $\underline{A}X = B$ has a solution $X \in \underline{X}$ for each bounded operator $B : \mathcal{X}_2 \to \mathcal{X}$.

In the following section we state the main results of this paper. The proofs will not be given here but may be found in [9].

4. MAIN RESULTS

With the application of Lemma 3.3 the following characterization of output regulation is obtained.

4.1 Proposition

If A_L is stable, output regulation is equivalent to the condition

$$\underline{C}_L \underline{A}_L^{-1} B_L = C_2$$

or, equivalently,

---

* Extended complex plane is the one-point compactification of the ordinary complex plane, by adjunction of the point ∞.

$$\begin{bmatrix} B_L \\ C_2 \end{bmatrix} \in \text{Im} \begin{bmatrix} A_{-L} \\ C_{-L} \end{bmatrix}.$$

In what follows, the composite system consisting of the original system and the exosystem will be considered as defined in the product space $X_1 \times X_2$, denoted by X. Then for $x \in X$, we write $x = (x_1, x_2)$, $x_1 \in X_1$, $x_2 \in X_2$. Also let

$$A = \begin{bmatrix} A_1 & A_3 \\ 0 & A_2 \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix},$$

and

$$C = [C_1 \quad C_2].$$

The equations

$$\dot{x}(t) = Ax(t) + Bu(t), \quad t \geq 0$$

$$y(t) = Cx(t)$$

then describe the composite system.

Finally, we make an additional assumption that

$$\text{the pair } (C, A) \text{ is detectable.} \qquad (4)$$

The solution of the main problem is now stated.

4.2 Theorem

Under the assumption (4), there exists a synthesis for the system if and only if

$$\begin{bmatrix} A_3 \\ C_2 \end{bmatrix} \in \text{Im} \begin{bmatrix} A_{-1} & B_{-1} \\ C_{-1} & 0 \end{bmatrix}. \qquad (5)$$

The most common class of exogenous signals against which regulation is to be achieved consists of step functions corresponding to the case $A_2 \equiv 0$. It is of interest, therefore, to specialize the above result to this case. Further, this yields a simpler criterion which is easy to verify in applications. The result may be stated as

4.3 Corollary

When $A_2 \equiv 0$, the necessary and sufficient condition for regulation with internal stability is that

$$\text{Im} \begin{bmatrix} A_3 \\ C_2 \end{bmatrix} \subset \text{Im} \begin{bmatrix} A_1 & B_1 \\ C_1 & 0 \end{bmatrix}. \qquad (6)$$

An essential feature of any synthesis for the system (1) is contained in

4.4 Theorem

If Σ_c is a synthesis for the system (1), then there exists an injective[*] operator $V: X_2 \to X_c$ such that the following diagram commutes i.e., for all $x_2 \in X_2$, $Vx_2 \in \mathcal{D}(A_c)$ and $A_c V x_2 = V A_2 x_2$.

$$\begin{tikzcd}
\mathcal{D}(A_c) \arrow[r, "A_c"] & \mathcal{D}(A_c) \\
X_2 \arrow[u, hook, "V"] \arrow[r, "A_2"] & X_2 \arrow[u, hook, "V"']
\end{tikzcd}$$

The interpretation of Theorem 4.4 is that ImV is an A_c invariant subspace of X_c and the restriction $A_c|\mathrm{Im}V$ is isomorphic to A_2. When this is the case, we say A_c incorporates an internal model[§] of A_2.

The above theorem establishes that any synthesis for the evolution system (1) necessarily incorporates an internal model of the exosystem. This is an extension, in a limited sense, of the Internal Model Principle of [1] for a much wider class of linear systems. Here we have assumed feedback structure and not imposed the requirement of structural stability. In this paper we shall not discuss structural stability. Some results in this direction are obtained in [9] under additional assumptions on the system operators.

In the next section we apply the above general results to a class of hereditary systems to obtain concrete conditions in terms of the system parameters.

5. APPLICATIONS

The result of Corollary 4.3 may conveniently be applied to Time Delay Systems (TDS) to obtain concrete results in $\mathbb{R}^n$. To this end, consider the TDS:

$$\dot{z}_1(t) = A_0 z_1(t) + A_1 z_1(t-\alpha) + B_1 u(t) + A_3 x_2(t), \quad t \geq 0, \tag{7a}$$

$$y(t) = C_1 z_1(t) + C_2 x_2(t). \tag{7b}$$

Here $z_1(t) \in \mathbb{R}^n$, $U = \mathbb{R}^m$ and $Y = \mathbb{R}^p$. A_0, A_1, B_1, A_3 etc. are constant matrices of compatible size.

The exogenous signals $x_2(t)$ are now generated by

$$\dot{x}_2(t) = 0, \quad x_2(0) \in X_2 = \mathbb{R}^q . \tag{7c}$$

Transforming the equation (7a) into an evolution equation in the Hilbert space $\mathbb{R}^n \times L_2(-\alpha,0;\mathbb{R}^n)$ and applying the Corollary 4.3 we obtain

---

[*] An operator $T: X \to Y$ is injective if and only if $\mathrm{Ker}\, T = 0$.

[§] Observe here a slight departure from the original definition of internal model in [1] which refers to a reduplication in A_c of the dynamic structure of the exosystem. Such 'reduplication' plays no role in this paper.

5.1 Corollary

For the TDS with external signals given by (7c), there exists a synthesis if and only if

$$\operatorname{Im}\begin{bmatrix} A_3 \\ C_2 \end{bmatrix} \subset \operatorname{Im}\begin{bmatrix} A_0+A_1 & B_1 \\ C_1 & 0 \end{bmatrix}$$

or, equivalently,

$$\operatorname{Rank}\begin{bmatrix} A_0+A_1 & B_1 & A_3 \\ C_1 & 0 & C_2 \end{bmatrix} = \operatorname{Rank}\begin{bmatrix} A_0+A_1 & B_1 \\ C_1 & 0 \end{bmatrix}.$$

For the general case, when $A_2 \neq 0$, Theorem 4.2 may be applied; however, the resulting condition is not expressed in a simple form as above. An example of such an application to show how the conditions can still be verified is given in [9].

6. CONCLUSIONS

Some aspects of the regulator theory of finite dimensional linear multivariable systems are studied in this paper for evolution systems defined in Banach spaces. Conditions are obtained for the existence of a regulating controller . Under the (major) assumption of feedback, a version of the Internal Model Principle is verified to hold for a large class of linear systems. Application of the abstract results is made to a class of hereditary systems to obtain simple verifiable conditions.

The important question of structural stability is not considered in this paper. This aspect requires additional technical developments since the class of perturbations on the system operators must be chosen and the behavior of the infinitesimal generator under such perturbations must be studied carefully. In [9] some results are obtained in this direction.

7. ACKNOWLEDGEMENTS

The author wishes to acknowledge Professor W.M. Wonham for his encouragement and help during the time this research was carried out.

REFERENCES

[1] B.A. Francis & W.M. Wonham, "The Internal Model Principle for Linear Multivariable Regulators", Applied Mathematics and Optimization, Vol. 2, No. 2, 1975, pp. 170-194.

[2] W.M. Wonham, Linear Multivariable Control: A Geometric Approach, Lecture Notes in Economics and Mathematical Systems, Vol. 101, Springer-Verlag, New York, 1974.

[3] E.J. Davison, "The Feedforward and Feedback Control of a General Servomechanism Problem", Proc. of 11th Allerton Conf. on Circuit and System Theory, Univ. of Illinois, 1973, pp. 343-362.

[4] E.J. Davison & A. Goldenberg, "Robust Control of a General Servomechanism Problem: The Servo Compensator", Automatica, Vol. 11, 1975, pp. 461-471.

[5] J.B. Pearson, R.W. Shields & P.W. Saats, Jr., "Robust Solutions to Linear Multivariable Control Problems", IEEE Tr. on Automatic Control, Vol. AC-9, No. 5, 1974, pp. 508-517.

[6] T. Kato, Perturbation Theory of Linear Operators, Springer-Verlag, New York, 1966.

[7] S.G. Krein, Linear Differential Equations in Banach Spaces, Translation of Mathematical Monographs, Vol. 29, AMS, Providence, R.I., 1971.

[8] A.V. Balakrishnan, "Optimal Control Problems in Banach Spaces", SIAM J. on Control, Ser. A, Vol. 3, No. 1, 1965, pp. 152-180.

[9] K.P.M. Bhat, "Regulator Theory for Evolution Systems", Ph.D. Thesis, in preparation, Dept. of Electrical Engineering, Univ. of Toronto, Toronto.

OPTIMUM NOTCH FILTERING FOR PULSED COMMUNICATION SIGNALS

Richard W. Koralek
San Jose State University
San Jose, California

and

Armin C. W. Reese
E S L, Inc.
Sunnyvale, California

Communication signals are often corrupted by single tone or narrowband interference. When this interference is filtered out by a notch filter, the notch causes intersymbol interference, degrading the original signal. The problem is to balance the amount of signal degradation due to the notch and the amount of interferer removed.

Consider a pulsed communication signal. A measure of the degradation introduced by the notch is the mean squared error. A more meaningful measure, however, is the mean squared error measured only during the signal on-times. It is shown, theoretically and experimentally, that for a notch filter of any given shape, the degradation is proportional to the area under the notch amplitude curve.

From this result, it can be seen that the only way to reduce signal degradation while still effectively removing the offending interferer is by adjusting the shape of the notch. A particular notch shape is shown that does reduce signal degradation by forcing most of the intersymbol interference energy to occur during the signal off-times. This notch shape can therefore be tailored to give minimum signal degradation for any given pulsed communication signal, and experimental results show that this notch gives significant improvement over a standard notch shape.

Some of the implications of these results are interesting. In particular, many adaptive notch filtering techniques seek to minimize the noise measured in the pulse off-times. However, these results would indicate that a properly shaped notch might actually maximize the energy in the pulse off-times under certain conditions. Further work is presently being done in the areas of other notch shapes, the adaptive filtering problem, and the application of these results to continuous communication signals.

NOTE: This study was performed at E S L, Inc., Sunnyvale, California during the summer of 1975.

NEW BIQUADS USING TWO OPERATIONAL AMPLIFIERS

KEIJI SUZUKI
Transmission Division, Nippon Electric Company, Ltd.
Kawasaki, Japan

ABSTRACT

This paper describes two new active RC filters capable of realizing any kind of biquadratic transfer function. The circuits apply an advanced analog computer realization method and are constructed with only two operational amplifiers and two capacitors. Design formulas, including the free parameter decision method, are discussed. It is shown that circuit sensitivity to passive components is very low and the operational amplifiers finite bandwidth influences were investigated. Using the proposed methods, two kinds of filters were designed and built. Experimental results show good performances.

I. INTRODUCTION

Cost and performance are the most important factors in active RC filters. Cost depends on the ease of realizing any biquadratic transfer function, the number of elements, tuning ease and so on. Performance depends on sensitivity, i.e., the variation of the transfer function due to changes in passive and active components.

Many circuit configurations have been presented up to now to satisfy cost and performance requirements. Two methods – the analog computer realization method [1] and the all pass loop method [2] – have satisfied many of the requirements. However, these circuits need three or four operational amplifiers to realize the general biquadratic transfer function. Therefore, they have a problem involving the number of active elements, which affects to their cost and power dissipation.

To overcome this problem, Soderstrand [3] and Nossek & Götz [4] proposed new circuits using only two operational amplifiers based on the analog computer realization method. However, the former circuit needs an additional operational amplifier to realize the finite transmission zero. The latter circuit is not capable of realizing the following numerator polynomial of the transfer function:

$$F(s) = s^2 + as + b \quad \text{where } b < 0. \tag{1}$$

In this paper, two new circuits are proposed, which is based on the analog computer realization method using only two operational amplifiers and two capacitors. The circuits are capable of realizing any kind of transfer function. Although the circuits use fewer operational amplifiers, they retain the advantages of the analog computer realization method.

New circuit configuration is discussed first. Then, design formulas, sensitivity to passive components and operational amplifier finite bandwidth influence are briefly described. The last section of the paper describes design examples and their experimental results.

II. NEW CIRCUIT CONFIGURATION

Consider the circuit shown in Fig. 1, where V_1 represents input voltage and the output voltage is taken from either V_2 or V_3. Assuming operational amplifiers to be ideal, the voltage transfer functions are readily obtained as follows:

$$T_1(s) = V_2/V_1 = F_1(s)/G(s) \tag{2}$$

$$T_2(s) = V_3/V_1 = F_2(s)/G(s) \tag{3}$$

where

$$
\begin{aligned}
G(s) &= (Y_4+Y_5)[Y_6Y_7(Y_{10}+Y_{11}+Y_{12})+Y_3Y_{12}(Y_7+Y_8+Y_9)] \\
F_1(s) &= Y_{12}(Y_7+Y_8+Y_9)[Y_4(Y_2+Y_3+Y_6)-Y_1Y_5]+Y_6(Y_4+Y_5)[Y_{10}(Y_7+Y_9)-Y_8(Y_{11}+Y_{12})] \qquad (4) \\
F_2(s) &= Y_7(Y_{10}+Y_{11}+Y_{12})[Y_4(Y_2+Y_3+Y_6)-Y_1Y_5]+Y_3(Y_4+Y_5)[Y_8(Y_{11}+Y_{12})-Y_{10}(Y_7+Y_9)].
\end{aligned}
$$

Admittances Y_i $(i = 1 - 12)$ can be selected so as to obtain the following biquadratic transfer function:

$$T(s) = (\alpha s^2+\beta s+\gamma)/(s^2+as+b). \qquad (5)$$

In this paper, two circuits, i.e., A and B, are considered.

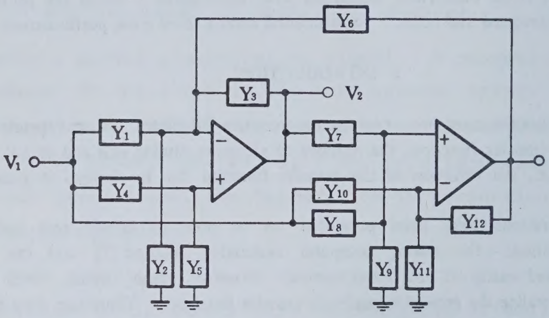


Fig. 1 New Circuit Configureation

A. Circuit A

In this circuit, the output voltage is taken from V_3, and Y_i are:

$$
\begin{aligned}
Y_3 &= sC_3 \\
Y_{12} &= sC_{12} \qquad (6) \\
Y_i(i \neq 3, 12) &= g_i
\end{aligned}
$$

In this case, the voltage transfer function is obtained as follows:

$$T_1(s) = V_3/V_1 = F_1(s)/G_1(s) \qquad (7)$$

where

$$
\begin{aligned}
G_1(s) &= s^2+g_6g_7s/C_3(g_7+g_8+g_9)+g_6g_7(g_{10}+g_{11})/C_3C_{12}(g_7+g_8+g_9) \\
F_1(s) &= g_4s^2/(g_4+g_5)+s/C_3\cdot[g_4(g_2+g_6)-g_1g_5/(g_4+g_5)-g_6g_8/(g_7+g_8+g_9)] \\
&\quad +g_6/C_3C_{12}(g_7+g_8+g_9)\cdot[g_{10}(g_7+g_9)-g_8g_{11}].
\end{aligned}
$$

Any biquadratic transfer function is obtained by choosing proper g_i's in Eq. (7). By comparing Eq. (5) with Eq. (7), circuits named A1 to A5 in Table 1 can be obtained.

Circuit No.	g_1	g_2	g_4	g_5	g_7	g_8	g_9	g_{10}	g_{11}	Transfer Function	Remarks
A1	0	0	0	#	#	0	*		*	$T_1(s)=\dfrac{g_6g_{10}/C_3C_{12}}{s^2+g_6g_7s/C_3(g_7+g_9)+g_6g_7(g_{10}+g_{11})/C_3C_{12}(g_7+g_9)}$	LPF
A2		0	0	#	#	0	*	0		$T_1(s)=\dfrac{-g_1s/C_3}{s^2+g_6g_7s/C_3(g_7+g_9)+g_6g_7g_{11}/C_3C_{12}(g_7+g_9)}$	BPF
A3		*			#	0	*	0		$T_1(s)=\dfrac{g_4s^2/(g_4+g_5)}{s^2+g_6g_7s/C_3(g_7+g_9)+g_6g_7g_{11}/C_3C_{12}(g_7+g_9)}$	HPF
A4		*			#	0	*			$T_1(s)=\dfrac{g_4s^2/(g_4+g_5)+g_6g_{10}/C_3C_{12}}{s^2+g_6g_7s/C_3(g_7+g_9)+g_6g_7(g_{10}+g_{11})/C_3C_{12}(g_7+g_9)}$	Notch
A5							*			$T_1(s)=\dfrac{F(s)}{s^2+g_6g_7s/C_3(g_7+g_8+g_9)+g_6g_7(g_{10}+g_{11})/C_3C_{12}(g_7+g_8+g_9)}$ where $F(s)=g_4s^2/(g_4+g_5)+s/C_3\cdot[g_4(g_2+g_6)/(g_4+g_5)-g_1g_5/(g_4+g_5)$ $-g_6g_8/(g_7+g_8+g_9)]+g_6/C_3C_{12}(g_7+g_8+g_9)\cdot[(g_7+g_9)g_{10}-g_8g_{11}]$	Complex −Zero +

Table 1. Circuit A Element Identification for Realizing Biquadratic Transfer Function

Notes

(1) $Y_3=sC_3,\ Y_{12}=sC_{12}$
(2) * These elements can be set to zero.
(3) # These elements can be set to infinite.
(4) + For all-pass $g_8=g_9=g_{11}=0$.

B. Circuit B

In this circuit, the output voltage is taken from V_3, and Y_i are:

$$Y_3=sC_3$$
$$Y_9=sC_9+g_9 \tag{8}$$
$$Y_i(i\neq 3,9)=g_i$$

In this case, the voltage transfer function is obtained as follows:

$$T_2(s)=V_3/V_1=F_2(s)/G_2(s) \tag{9}$$

where

$$G_2(s)=s^2+(g_7+g_8+g_9)s/C_9+g_6g_7(g_{10}+g_{11}+g_{12})/C_3C_9g_{12}$$

$$F_2(s)=-g_{10}s^2/g_{12}-s/C_9g_{12}\cdot[g_9g_{10}+g_5g_7g_{10}/(g_4+g_5)-g_8(g_{11}+g_{12})-g_4g_7(g_{11}+g_{12})/(g_4+g_5)]$$
$$-g_7(g_{10}+g_{11}+g_{12})/C_3C_9g_{12}(g_4+g_5)\cdot[g_1g_5-g_4(g_2+g_6)]$$

By choosing proper g_i's, circuits named B1 to B5 in Table 2 can be also obtained.

Circuit No.	g_1	g_2	g_4	g_5	g_8	g_9	g_{10}	g_{11}	g_{12}	Transfer Function	Remarks
B1		0	0	#	0	*	0	*	#	$T_2(s)=\dfrac{-g_1g_7(g_{11}+g_{12})/C_3C_9g_{12}}{s^2+(g_7+g_9)s/C_9+g_6g_7(g_{11}+g_{12})/C_3C_9g_{12}}$	LPF
B2	0	0	0	#		*	0	*	#	$T_2(s)=\dfrac{g_8(g_{11}+g_{12})s/C_9g_{12}}{s^2+(g_7+g_8+g_9)s/C_9+g_6g_7(g_{11}+g_{12})/C_3C_9g_{12}}$	BPF
B3	0	0	0	#		*		*		$T_2(s)=\dfrac{-g_{10}s/g_{12}}{s^2+(g_7+g_8+g_9)s/C_9+g_6g_7(g_{10}+g_{11}+g_{12})/C_3C_9g_{12}}$	HPF
B4		0	0	#		*		*		$T_2(s)=\dfrac{-g_{10}s^2/g_{12}-g_1g_7(g_{10}+g_{11}+g_{12})/C_3C_9g_{12}}{s^2+(g_7+g_8+g_9)s/C_9+g_6g_7(g_{10}+g_{11}+g_{12})/C_3C_9g_{12}}$	Notch
B5		*				*				$T_2(s)=\dfrac{-F(s)}{s^2+(g_7+g_8+g_9)s/C_9+g_6g_7(g_{10}+g_{11}+g_{12})/C_3C_9g_{12}}$ where $F(s)=g_{10}s^2/g_{12}$ $+s/C_9\cdot[g_5g_{10}(g_7+g_9)/g_{12}(g_4+g_5)+g_4g_9g_{10}/g_{12}(g_4+g_5)$ $-g_8(g_{11}+g_{12})/g_{12}-g_4g_7(g_{11}+g_{12})/g_{12}(g_4+g_5)]$ $+g_7(g_{10}+g_{11}+g_{12})/g_{12}(g_4+g_5)\cdot[g_1g_5-g_4(g_2+g_6)]$	Complex −Zero +

Table 2. Circuit B Element Identification for Realizing Biquadratic Transfer Function

Notes

(1) $Y_6=g_6$, $Y_7=g_7$ always

(2) $Y_3=sC_3$, $Y_9=sC_9+g_9$

(3) + For all-pass $g_2=g_4=0$.

III. DESIGN PROCEDURES

Design formulas for the circuits A and B are shown in Table 3.

Realized Transfer Function	Design Formulas	
	Circuit A	Circuit B
$\dfrac{\gamma}{s^2+as+b}$	$l\equiv g_7/(g_7+g_9)$, $g_6=aC_3/l$, $g_{10}=l\gamma C_{12}/a$, $g_{11}=C_{12}/a\cdot(b-l\gamma)$, $b\geqq l\gamma$, $l\leqq 1$	$n\equiv g_{12}/(g_{11}+g_{12})$, $h\equiv g_9/g_7$, $g_7=aC_9/(1+h)$ $g_6=nb(1+h)C_3/a$, $g_1=n\gamma(1+h)C_3/a$, $n\leqq 1$
$\dfrac{\beta s}{s^2+as+b}$	$l\equiv g_7/(g_7+g_9)$, $g_6=aC_3/l$, $g_1=\beta C_3$, $g_{11}=bC_{12}/a$, $l\leqq 1$	$n\equiv g_{12}/(g_{11}+g_{12})$, $h\equiv g_9/g_7$, $g_8=n\beta C_9$, $g_7=(a-n\beta)C_9/(1+h)$, $g_6=nb(1+h)C_3/(a-n\beta)$, $a>n\beta$, $n\leqq 1$
$\dfrac{\alpha s^2}{s^2+as+b}$	$l\equiv g_7/(g_7+g_9)$, $h\equiv g_2/g_1$, $g_6=aC_3/l$, $g_{11}=bC_{12}/a$, $g_1=a\alpha C_3/l[1-\alpha(1+h)]$, $g_5/g_4=(1-\alpha)/\alpha$, $l\leqq 1$, $\alpha(1+h)<1$	$n\equiv g_{12}/(g_{11}+g_{12})$, $h\equiv g_9/g_7$, $g_8=an\alpha C_9/(n\alpha+1)$, $g_9=aC_9/(1+h)(n\alpha+1)$, $g_6=nb(1+h)C_3/a$, $g_{10}/g_{12}=\alpha$, $n\leqq 1$
$\dfrac{\alpha s^2+\gamma}{s^2+as+b}$	$l\equiv g_7/(g_7+g_9)$, $h\equiv g_2/g_1$, $g_6=aC_3/l$, $g_{10}=l\gamma C_{12}/a$, $g_{11}=(b-\gamma l)C_{12}/a$, $g_5/g_4=(1-\alpha)/\alpha$, $g_1=a\alpha C_3/l[1-\alpha(1+h)]$, $h\geqq l\gamma$, $\alpha(1+h)<1$, $l\leqq 1$	$n\equiv g_2/(g_{11}+g_{12})$, $h\equiv g_9/g_7$, $g_8=an\alpha C_9/(n\alpha+1)$, $g_7=aC_9/(1+h)(n\alpha+1)$, $g_6=nb(1+h)C_3/a$, $g_1=n\gamma(1+h)C_3/a$, $g_{10}/g_{12}=\alpha$, $n\leqq 1$
$\dfrac{\alpha s^2+\beta s+\gamma}{s^2+as+b}$	$l\equiv g_7/(g_7+g_8+g_9)$, $m\equiv g_8/(g_7+g_8+g_9)$, $h\equiv g_2/g_1$, $g_6=aC_3/l$, $g_{10}=lC_{12}/a\cdot(mb/l+\gamma)$, $g_{11}=C_{12}/a\cdot[b(1-m)-l\gamma]$, $g_5/g_4=(1-\alpha)/\alpha$, $g_1=C_3[a(\alpha-m)-\beta l]/l[1-\alpha(1+h)]$, $g_8/g_7=m/l$, $g_9/g_7=(1-l-m)/l$	$l\equiv g_5/(g_4+g_5)$, $n\equiv g_{12}/(g_{11}+g_{12})$, $h\equiv g_9/g_7$, $q\equiv g_2/g_1$, $g_7=(a+n\beta)C_9/(l+h)(n\alpha+1)$, $g_6=nb(l+h)C_3/(n\beta+a)$, $g_{10}/g_{12}=\alpha$, $g_8=C_9[a-(1+h)(n\beta+a)/(l+h)(n\alpha+1)]$, $g_1=nC_3(l+h)[\gamma+b(1-l)]/(n\beta+a)[l-(1-l)q]$

Table 3. Design Formulas

Several degrees of freedom exist in choosing element values, as shown in Table 3. These can be used to obtain low sensitivity or low element-value-spread or maximum dynamic range. In case of a high Q circuit, the most important parameter in these circuits is the dynamic range. Because if free parameters are arbitarary selected, a large difference between the amplifier's output voltages might occur and the amplifiers would be overdriven. Therefore, taking into account the output voltage of each amplifier, some free parameters should be determined. Now, consider the transfer function expressed in Eq. (5). In case of a high Q circuit, the maximum output voltage of each amplifier occurs close to $s=i\sqrt{b}$. Therefore, the following parameter k is considered.

$$k=|T_2(j\sqrt{b})|/|T_1(j\sqrt{b})|=|V_3(j\sqrt{b})|/|V_2(j\sqrt{b})| \qquad (10)$$

k is selected so as to obtain the maximum dynamic range. Decision formulas for some free parameters in circuits A and B are shown in Table 4.

Realized Transfer Function	Decision Method	
	Circuit A	Circuit B
$\frac{\gamma}{s^2+as+b}$	$l=k/Q_0$	$n=\sqrt{1/(1+Q^2)}/k(1+h)$
$\frac{\beta s}{s^2+as+b}$	$l=k/\sqrt{1+Q_0^2}$	$n=1/[kQ_0(1+h)+\beta/a]$
$\frac{\alpha s^2}{s^2+as+b}$	$l=k/\sqrt{1+Q_0^2}$	$n=1/kQ_0(1+h)$
$\frac{\alpha s^2+\gamma}{s^2+as+b}$	$l=k\sqrt{\left(\frac{\gamma}{\alpha b}-1\right)^2 \Big/ \left[1+Q_0^2\left(1-\frac{\gamma}{\alpha b}\right)^2\right]}$	$n=\sqrt{\left(1-\frac{\alpha b}{\gamma}\right)^2 \Big/ \left[1+Q_0^2\left(\frac{\alpha b}{\gamma}-1\right)^2\right]} \Big/ k(1+h)$
$\alpha\frac{s^2-as+b}{s^2+as+b}$	$l=k/\sqrt{1+4Q_0^2}$	$n=1/[2kQ_0(1+h)+\alpha]$

Table 4. Some Free Parameters Decision Method

Notes

$$Q_0=\sqrt{b}/a$$

IV. SENSITIVITY

Assuming operational amplifiers to be ideal, consider Q_0– and ω_0–sensitivities to passive components. The sensitivity of a quantity P to an element x is given by:

$$S^P_x=x/P\cdot\partial P/\partial x \qquad (11)$$

Q_0 and ω_0 of circuits A and B are expressed as

$$\left.\begin{aligned} Q_0&=\sqrt{C_3(g_{10}+g_{11})(g_7+g_8+g_9)/C_{12}g_6g_7} \\ \omega_0&=\sqrt{g_6g_7(g_{10}+g_{11})/C_3C_{12}(g_7+g_3+g_9)} \end{aligned}\right\} \quad \underline{\text{circuit-A}} \qquad (12)$$

$$\left.\begin{aligned} Q_0&=1/(g_7+g_8+g_9)\sqrt{C_9g_6g_7(g_{10}+g_{11}+g_{12})/C_3g_{12}} \\ \omega_0&=\sqrt{g_6g_7(g_{10}+g_{11}+g_{12})/C_3C_9g_{12}} \end{aligned}\right\} \quad \underline{\text{circuit-B}} \qquad (13)$$

The use of Eqs. (11)-(13) leads to :

$$\left|S^P_x\right|\leqq 1/2 \qquad (14)$$

where P represents Q_0 or ω_0 and x represents any capacitance or conductance. The Q_0 sensitivity is independent of the realized Q.

V. OPERATIONAL AMPLIFIER FINITE BANDWIDTH INFLUENCE

The operational amplifier finite bandwidth limits the frequency range of the active RC filter. In this section, operational amplifier finite bandwidth influence on Q_0 and ω_0 is investigated. The desired characteristic equation is normalized as follows:

$$G(s)=s^2+1/Q_0\cdot s+1 \qquad (15)$$

Assuming the gain characteristic of operational amplifiers as $A_1(s)=A_2(s)=A(s)\simeq B/s$ (where B expresses the normalized gain-bandwidth product), the characteristic equation can be expressed as follows:

$$G^*(s)=G(s)+G_x(s)/A(s) \qquad (16)$$

To obtain pole-displacement Δs_0, the following result reported by Fleisher [5] is used:

$$\Delta s_0 \simeq -1/A(s_0)\cdot G_x(s_0)/(s_0-s_0^*) \qquad (17)$$

where s_0 and s_0^* express the dominant pole and its conjugate, respectively. To relate the variable quantities of ω_0 and Q_0 with the pole-displacement, the result reported by Moschytz [6] is also used:

$$\Delta s_0/s_0=\Delta\omega_0/\omega_0-j\Delta Q_0/Q_0\cdot[1/(4Q_0^2-1)]^{1/2}\simeq\Delta\omega_0/\omega_0-j1/2Q_0\cdot\Delta Q_0/Q_0 \qquad (18)$$

$\Delta\omega_0/\omega_0$, $\Delta Q_0/Q_0$ obtained by using Eqs. (18)-(19), are shown in Table 5.

Circuit No.	$\Delta\omega_0/\omega_0$		$\Delta Q_0/Q_0$	
	Circuit A	Circuit B	Circuit A	Circuit B
1	$-[Q_0+1/lQ_0-4/Q_0]/2B$	$-nQ_0(1+\gamma)(1+h)/2B$	$Q_0/2B\cdot(5+1/lQ_0^2-2/Q_0^2)$	$Q_0/2B\cdot[2(1+1/n)$ $-n(1+\gamma)(1+h)]$
2	$-[Q_0+\beta+1/lQ_0-4/Q_0]/2B$	$-nQ_0(1+h)/2B(1-n\beta Q_0)$	$Q_0/2B\cdot(5+\beta/Q_0+1/lQ_0^2-2/Q_0^2)$	$Q_0/2B\cdot[2(1+1/n)$ $-n(1+h)/(1-n\beta Q_0)]$
3	$-1/2B\cdot[Q_0+1/lQ_0\{1-\alpha(1+h)\}$ $-4/Q_0]$	$-nQ_0(1+h)/2B$	$Q_0/2B\cdot[5+1/lQ_0^2\{1-\alpha(1+h)\}$ $-2/Q_0^2]$	$Q_0/2B\cdot[2(1+n/(1+n\alpha))$ $-n(1+h)]$
4	$-1/2B\cdot[Q_0+1/lQ_0\{1-\alpha(1+h)\}$ $-4/Q_0]$	$-nQ_0(1+\gamma)(1+h)/2B$	$Q_0/2B\cdot[5+1/lQ_0^2\{1-\alpha(1+h)\}$ $-2/Q_0^2]$	$Q_0/2B\cdot[2(1+n/(1+n\alpha))$ $-n(1+\gamma)(1+h)]$
5	$-1/2B\cdot[Q_0-4/Q_0+\{1+\alpha h(l+1)\}$ $/lQ_0\{1-\alpha(1+h)\}]$	$-nQ_0(1+h)(1+\alpha/l)$ $/2B(1-n\alpha)$	$Q_0/2B\cdot[5-2/Q_0^2+\{1+\alpha h(l+1)\}$ $/lQ_0^2\{1-\alpha(1+h)\}]$	$Q_0/2B\cdot[2(1+n/(1+n\alpha))$ $-nQ_0(1+h)(1+\alpha/l)/(1-n\alpha)]$

Table 5. Finite Bandwidth Influence

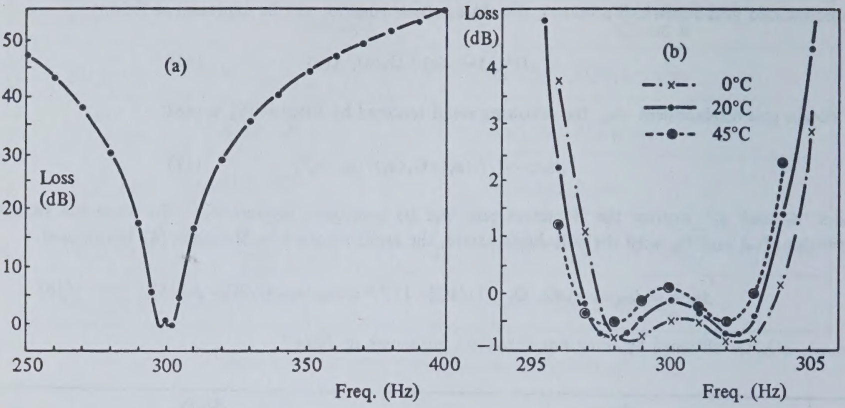
These values are the same as values of usual analog computer realization, using two operational amplifiers. If the nondominant pole influence is small, these variations can be compensated for by transfer fuction predistortion in the design stage.

VI. EXAMPLES

Two filters were designed and built by using the proposed circuits–A and B.

(1) A Fourth-Order High Q Band-Pass Filter (BPF)

This BPF has a 300 Hz center frequency, 0.5 dB maximum passband attenuation and 8.5 Hz bandwidth. Minimum stopband attenuation is 50 dB outside the frequency 380 Hz and Q of each section in 68. Resistors (temperature coefficient ≦±100 ppm/°C, tolerance ≦±1 percent), capacitors (temperature coefficient = 0 – 70 ppm/°C, tolerance ≦±1 percent) and a 741 type dual operational amplifier were used. The realization uses cascaded sections of circuit A2, but it is modified to reduce the resistor-value-spread by using a pi-to-tee transformation. The measured frequency response is shown in Fig. 2. A 0.5 Hz center frequency deviation from initial characteristic was measured in the 0°C to 45°C temperature range.



$$T(s)=\frac{0.0243s}{(s^2+0.0144s+0.96)}\cdot\frac{0.0243s}{(s^2+0.0147s+1.0)}\quad \text{(Normalized Freq. = 303 Hz)}$$

(a) Mesaured Frequency Response
(b) Passband Frequency Response at 0°C, 20°C and 45°C

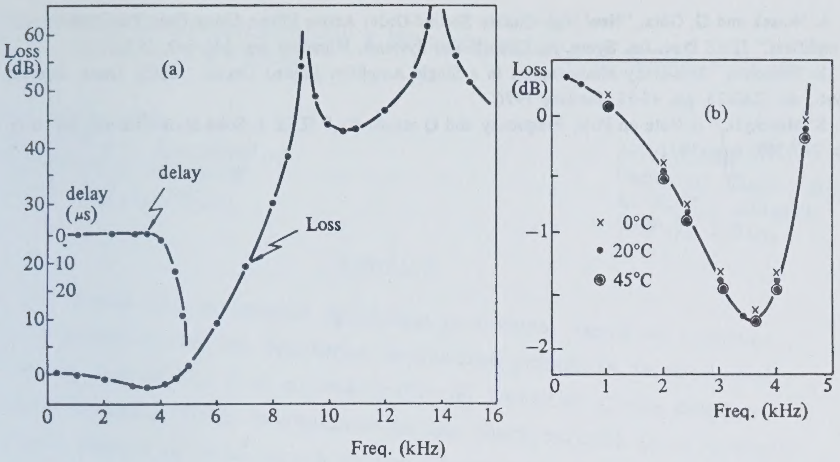
Fig. 2 Fourth-Order Band-Pass Filter

(2) An Eleventh-Order Shaping Low-Pass Filter (LPF)

This LPF has a passband attenuation characteristic expressed as

$$20\log_{10}\frac{\sin(\pi f/f_0)}{(\pi f/f_0)}\ (f_0=9600\ \text{Hz}) \qquad (19)$$

and maximally flat group delay in the passband, where the passband means 0 to 3 kHz frequency band. This filter is used to reconstruct an analog signal from amplitude modulated pulses. The realized circuit consists of cascaded sections of circuit B4-B5 and the first-order section. To compensate for the finite bandwidth influence, a transfer function predistortion is performed. Resistors (temperature coefficient ≧−350 ppm/°C, tolerance ≦±1 percent), capacitors (temperature coefficient ≧−200 ppm/°C, tolerance ≦±1 percent) and a 741 type dual operational amplifier were used. The measured frequency response is shown in Fig. 3. Although this circuit was not tuned, the frequency response deviation due to initial elements deviation and the 0 °C to 45 °C temperature range remained less than 0.3 dB in the passband. Group delay deviation was negligible in the passband.



$$T(s)=\frac{0.131(s^2+524.0)}{(s^2+8.62s+54.8)}\cdot\frac{1.59(s^2-20.6)}{(s^2+12.6s+41.9)}\cdot\frac{1.46(s^2-8.16s+33.9)}{(s^2+11.9s+44.0)}$$

$$\frac{1.37(s^2+8.16s+33.9)}{(s^2+10.6s+48.0)}\cdot\frac{0.278(s^2+225.0)}{(s^2+5.44s+67.2)}\cdot\frac{6.42}{(s+6.42)}$$

(Normalized Freq. = 600 Hz)

(a) Measured Frequency Response (Loss and Delay)
(b) Passband Frequency Response at 0°C, 20°C and 45°C

Fig. 3. Eleventh-Order Shaping Low-Pass Filter

VII. CONCLUSION

New configurations have been proposed for active RC filter synthesis with the following features:

1) Only two operational amplifiers and two capacitors are required for realizing arbitrary bequadratic transfer functions.
2) All biquadratic transfer functions can be realized by the same configuration.
   As existing method, new method has also the following features:
3) These new configurations can be cascaded without isolating amplifiers.
4) Sensitivity to passive elements is very small.
   And finite bandwidth influence is the same as usual analog computer realization, using two operational amplifiers.

VIII. ACKNOWLEDGEMENT

The author wishes to thank M. Takahashi and T. Uchida for their support and E. Inamasu for his assistance in the course of this work.

IX. REFERENCES

(1) J. Tow, "Active RC Filters—A State Space Realization," Proc. IEEE, vol. 56, pp. 1137-1139, June 1968
(2) R. Tarmy and M. S. Ghausi, "Very High-Q Insensitive Active RC Networks," IEEE Trans. Circuit Theory, vol. CT-17, pp. 358-366, August 1970
(3) M. A. Soderstrand, "Biquads Using Only Two Operational Amplifiers," Proc. 1974 Int. Symo. on Circuits and Systems, San Francisco pp. 651-655

(4) J. A. Nossek and G. Götz, "New High-Quality Second-Order Active Filters Using Only Two Operational Amplifiers," IEEE Proc. Int. Symp. on Circuits and Systems, München, pp. 446-449, 1976

(5) P. E. Fleischer, "Sensitivity Minimization in a Single Amplifier Biquad Circuit," IEEE Trans. Circuits Syst., vol. CAS-23, pp. 45-55, January 1976

(6) G. S. Moschytz, "A Note on Pole, Frequency and Q sensitivity," IEEE J. Solid-State Circuits, vol. sc-6, pp. 267-269, Aug. 1971

ACTIVE RC MULTIPORT SYNTHESIS OF A TRANSFER ADMITTANCE MATRIX WITH A MINIMUM NUMBER OF CAPACITORS.

KHALIL AHMAD KHAN
Dept. of Elect. Engineering
ALIGARH MUSLIM UNIVERSITY
ALIGARH, (U.P.), India.

AFAQ AHMAD
Dept. of Elect. Engg.
A. M. U., Aligarh
(U.P.), India.

ABSTRACT

A simple and systematic synthesis procedure, based on a state-variable approach and the reactance extraction principle is presented, whereby any qxp matrix T(s) of real rational function of the complex frequency variable s can be realized as the short-circuit (s.c.)transfer admittance matrix of a multiport active RC network. The realization requires a minimum number of n-capacitors with unity capacitance spread, n equal to the degree of T(s) and at the most (p+2n) inverting grounded voltage amplifiers. All the capacitors and ports will share a common ground.

INTRODUCTION

With the advent of integrated circuit technology, a great deal of interest has been evinced in active RC synthesis of multiport network functions with a minimum number of capacitors during recent years [1]-[5]. Bickart and Melvin [1], based on their earlier paper [2], gave an interesting synthesis procedure for realizing various multiport network functions with grounded ports. Recently Lal and Khan [3] - [4] suggested a simplified method for synthesizing a pxp s.c. admittance matrix and a qxp open-circuit (o.c.) transfer impedance matrix using operational amplifiers as active elements respectively. The methods are essentially based on a state-variable technique and the reactance extraction principle. The main advantage of their method [3] or [4] over [1] is that the upper bound on the number of active elements is reduced from 2(p+n) to (p+2n) only and it can be reasonably expected to require fewer resistors while retaining all other advantages of [1].

This paper presents a simple and systematic synthesis procedure, similar to [3] - [4], whereby any qxp matrix T(s) of real rational functions of the complex frequency variable s, can be realized as the s.c. transfer admittance matrix of a (q+p) port active RC network. The realization requires a minimum number of n grounded capacitors with unity capacitance spread, n equal to the degree of T(s) and at the most (p+2n) inverting grounded voltage amplifiers. The structure of the circuit in terms of the minimum number of elements and grounded ports makes it particularly suitable for integrated circuit fabrication.

SYNTHESIS PROCEDURE

In the synthesis procedure to be presented, we take T(s) to be regular at $s=\infty$; if it is not so, a Möbius transformation [6] can always be invoked to remove singularities at $s=\infty$. Then the synthesis procedure is applied to the transformed matrix, after which the capacitors in the network are each replaced by a capacitor and a resistor in series to get a realization for T(s). The general idea in the proposed approach is to realize a given multiport network function T(s), with the help of capacitive and resistive networks and, by introducing suitable active elements, to force the s.c. conductance matrix ($\bar{G}$) of the resistive network to be hyperdominant which can be easily realized, while ensuring that the state-model of the realized network corresponds to the state model obtained from the given network function.

A minimal realization set $\{A,B,C,D\}$ associated with T(s) can be easily constructed by applying any standard algorithm [7] to give the state equations of the form

$$\dot{e}_2 = Ae_2 + Be_1 \tag{1a}$$

$$i_3 = Ce_2 + De_1 \tag{1b}$$

where e_1 and i_3 are p-vector of excitation port voltages and q-vectors of response port currents respectively and e_2, an n-vector of state-variables, is the port voltages at the ports common to N_R and N_C subnetworks as shown in Fig. 1a. $\{A,B,C,D\}$ are real constant matrices, A, having the minimum dimension $n = \text{degree}\,\{T(s)\}$.

The relationship imposed by N_C on e_2 and i_2 is

$$i_2 = -\mathscr{C}\dot{e}_2 \tag{2}$$

where $\mathscr{C}$ is an nxn non-singular matrix and can be assumed to be diagonal matrix with positive entries only, resulting in a capacitive subnetwork in the form of a star of n-capacitors, thereby ensuring that no more than n capacitors are needed in the realization. That the realization requires at least n capacitors follows from the fact that $\mathscr{C}$ would be singular if the subnetwork N_C contained fewer than n-capacitors.

Assuming the subnetwork, N_R, consisting of resistors and inverting grounded voltage amplifiers to have a structure shown in Fig. 1b. Where $\bar{N}_R$ is a (2p+q+3n) port common terminal resistive network.

The s.c. parameter equation of $\overline{N}_R$ can be written as

$$I = \overline{G}E \tag{3}$$

where

$$I = [i_1, i_2, i_3, i_4, i_5, i_6]',$$

$$E = [e_1, e_2, e_3, e_4, e_5, e_6]'$$

and

$$\overline{G} = [g_{ij}]_{(2p+q+3n)x(2p+q+3n)}$$

where i_1, i_4, e_1 and e_4 are each p-vectors, i_3 and e_3 are each q-vectors and i_2, i_5, i_6, e_2, e_5 and e_6 are each n-vectors; and the elements of $\overline{G}$ are the sub-matrices with $g_{ij}=g_{ji}'$, where prime denotes matrix transposition. $\overline{G}$ being the s.c. conductance matrix of a common ground resistive network has to be hyperdominant, a necessary and sufficient condition for the (2p+q+3n) port common terminal resistive $\overline{N}_R$ to be realizable without internal nodes [7]

The following constraints are imposed due to active elements (Fig. 1b)

$$\left.\begin{aligned} e_4 &= Ke_1 \\ e_5 &= Qe_2 \\ e_6 &= HQe_2 \end{aligned}\right\} \tag{4}$$

where the matrices K, Q and H are

$K = \text{diag.}\{k_1, \text{----}, k_p\}$, with $k_i \leqslant 0$ for all i,

$Q = \text{diag.}\{q_1, \text{----}, q_n\}$, with $q_j \leqslant 0$ for all j ,

and

$H = \text{diag.}\{h_1, \text{----}, h_n\}$, with $h_j \leqslant 0$ for all j .

From (2), (3) and (4), the state equations of the form (1) can be established and $\{A,B,C,D\}$ can be expressed as

$$A = -\mathcal{C}^{-1}(g_{22}+g_{25}Q+g_{26}HQ) \tag{5a}$$

$$B = -\mathcal{C}^{-1}(g_{21}+g_{24}K) \tag{5b}$$

$$C = (g_{32}+g_{35}Q+g_{36}HQ) \tag{5c}$$

$$D = (g_{31}+g_{34}K) \tag{5d}$$

Thus the problem of realization of T(s) has been reduced to that of specifying the g-sub-matrices, associated with $\overline{N}_R$ and given by (5), subject to the condition that $\overline{G}$ is hyper dominant. The existence of such a realization is evident from the following steps in the procedure for specifying the various sub-matrices of $\overline{G}$

Step I: Since D is the sum of non-positive and non-negative matrices, (5d) can be written as,

$$D = P + M$$

where

$P = \left[p_{ij}\right]_{pxn}$ contains all the non-negative entries of $(g_{31}+g_{34}K)$

Therefore,

$$g_{34}K = P$$

$$g_{31} = M.$$

Selecting suitable values for K; the submatrices g_{31} and g_{34} can be easily found out such that these have all non-positive entries.

Step II: Write down the (5c) as

$$C = P + M$$

then, obviously

$$g_{35} Q = P$$

and

$$g_{32} + g_{36} HQ = M.$$

Choosing suitable values for Q, H and g_{32} it is easy to get g_{35} and g_{36}. It is obvious that, by making $|K_i|$, $|q_j|$ and $|h_j|$ sufficiently large, the non zero elements of g_{34}, g_{35} and g_{36} can be made as small in magnitude as is necessary to make the rows of $[g_{31}, \ldots\ldots\ldots\ldots, g_{36}]$ hyperdominant. It can be easily shown that it is always possible to select such amplifier gains.

Step III: Selecting a suitable value for $\wp$, we can easily determine the sub-matrices g_{21} and g_{24} from (5b), such that these have all non-positive entries.

Step IV: From (5a), we have

$$g_{22} + g_{25}Q + g_{26} HQ = -\wp \underline{A} = P + M$$

where

$$g_{22} + g_{25}Q = P$$

and

$$g_{26} HQ = M.$$

Since $\wp$, Q and H are already known, then selecting a suitable value for the hyperdominant submatrix g_{22}, the submatrices g_{25} and g_{26} can be easily obtained such that these have non positive entries. We may also select g_{22} in such a way that the modulus of the rows corresponding to it is just equal to zero, thereby reducing the number of resistors required in the realization.

Thus having determined the g's that appear in (5), the remaining entries of $\overline{G}$ can be filled in arbitrarily; however, a maximum number of zero entries such that the hyperdominant nature of $\bar{G}$ is retained, is advantageous.

EXAMPLE

To illustrate the above results, the following 2x2 matrix T(s) will be realized as a s.c. transfer admittance matrix of a two port:

$$T(s) = \begin{bmatrix} s & -s \\ -s & s \end{bmatrix},$$

since T(s) is not regular at s = ∞, a Möbius transformation is needed,

Let $s = \frac{z}{1-z}$,

Then $$T(z) = \begin{bmatrix} \frac{z}{1-z} & \frac{-z}{1-z} \\ \frac{-z}{1-z} & \frac{z}{1-z} \end{bmatrix},$$

Obviously,

$$D = T(z)\Big|_{z=\infty} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Here p = q = 2; and n = 1

Now a minimal realization set $\{A, B, C\}$ can be obtained by using Chen and Mittal algorithm [7] to the transformed matrix function (T (z)-D).

Thus

$$A = [1] \;;\; B=[1 \;\, -1] \;;\; \text{and } C = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Choose

$$K = \begin{bmatrix} -10 & 0 \\ 0 & -10 \end{bmatrix}$$

From (5d), g_{31} and g_{34} can be obtained as

$$g_{31} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}; \text{ and}$$

$$g_{34} = \begin{bmatrix} 0 & -1/10 \\ -1/10 & 0 \end{bmatrix}.$$

Select

$$Q = [-10] \;;\; H = [-10] \text{ and } g_{32} = [0]_{2x1}$$

then from (5c)

$$g_{35} = \begin{bmatrix} 0 \\ -1/10 \end{bmatrix}$$

and

$$g_{36} = \begin{bmatrix} -1/100 \\ 0 \end{bmatrix}$$

Choose

$$\varphi = [1] \ ,$$

From (5b), g_{21} and g_{24} can be found out as

$$g_{21} = [-1 \quad 0], \text{ and}$$

$$g_{24} = [0 \quad -1/10] \ .$$

Select

$$g_{22} = [2]$$

From (5a), we obtain

$$g_{25} = [0] \ , \text{ and}$$

$$g_{26} = [-3/100] \ .$$

The remaining entries of $\bar{G}$ can be filled in arbitrarily such that it remains hyperdominant. A suitable choice of $\bar{G}$ is given below.

$$\bar{G} = \begin{bmatrix} 2 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 2 & 0 & 0 & 0 & -1/10 & 0 & -3/100 \\ -1 & 0 & 0 & 111/100 & 0 & 0 & -1/10 & 0 & -1/100 \\ 0 & -1 & 0 & 0 & 6/5 & -1/10 & 0 & -1/10 & 0 \\ 0 & 0 & 0 & 0 & -1/10 & 1/10 & 0 & 0 & 0 \\ 0 & 0 & -1/10 & -1/10 & 0 & 0 & 1/5 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1/10 & 0 & 0 & 1/10 & 0 \\ 0 & 0 & -3/100 & -1/100 & 0 & 0 & 0 & 0 & 1/25 \end{bmatrix}$$

Having obtained the $\bar{G}$-matrix of $\bar{N}_R$, the network can now be constructed as shown in Fig. 2. The number of active elements used in the realization is four i.e. (p+2n) instead (2p+2n) **required in [2]** . It may be noted that a capacitor and a resistor are in series at each capacitive port of $\bar{N}_R$ as Möbius transformation has been invoked in this example. Further, the network $\bar{N}_R$ (Figure 2) which realizes the hyperdominant $\bar{G}$ has no internal nodes as indicated earlier.

CONCLUSIONS

A simplified synthesis procedure has been developed whereby any qxp matrix T(s) of real rational functions of the complex frequency variable s can be realized as the s.c. transfer admittance matrix of an active RC multiport (filter) network with the ports grounded. The realized network requires a minimum number of n grounded capacitors having unity capacitance spread, n being the degree of T(s), and at the most (p+2n) grounded inverting voltage amplifiers which can be easily constructed with the help of the commercially available operational amplifiers. The facts that all the minimum number of capacitors are of equal value and that all the active elements, capacitors and ports are grounded are very much desirable if the network is to be fabricated as an integrated circuit.

Since a minimum number of capacitors are used, it is conjectured that the realization of T(s) will be relatively insensitive to capacitance variations. It is hoped that further investigations of this synthesis procedure will provide a quantitative assessment of the sensitivity of selected network attributes and validity of the conjecture. Finally, the procedure can be reduced rather easily to a digital computer programme.

REFERENCES

[1] T. A. Bickart and D. W. Melvin, "Synthesis of active RC multiport networks with grounded ports," Journ. of Franklin Institute, Vol. 294, No.4, pp. 289-312, November 1972.

[2] D. W. Melvin and T. A. Bickart, "p-port active RC network: short-circuit admittance matrix synthesis with a minimum number of capacitors," IEEE Trans. on Circuit Theory, Vol. CT-18, No.6, pp. 587-592, November 1971.

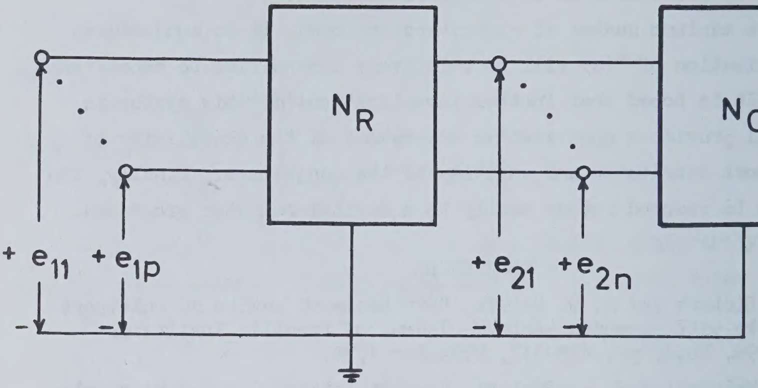
[3] M. Lal and K. A. Khan, "On active RC multiport network synthesis of short-circuit admittance matrix with a minimum number of capacitors," Presented at the IEEE INT. SYMP. on circuits and system Theory, held at San Francisco, U.S.A., April 21-25, 1974.

[4] K. A. Khan and M. Lal, "Active RC multipart synthesis of a transfer im-pedance matrix with operational Amplifiers," National system conference, held at IIT Delhi (INDIA), November 8-10, 1974.

[5] P. A. Ramamoorthy, K. Thulasiraman and K. Sankara Rao, "Synthesis of rational voltage transfer matrices using minimum number of capacitors with operational amplifiers," Proc. IEEE. INT. SYMP. on Circuit Theory, pp. 294-296, April, 1973.

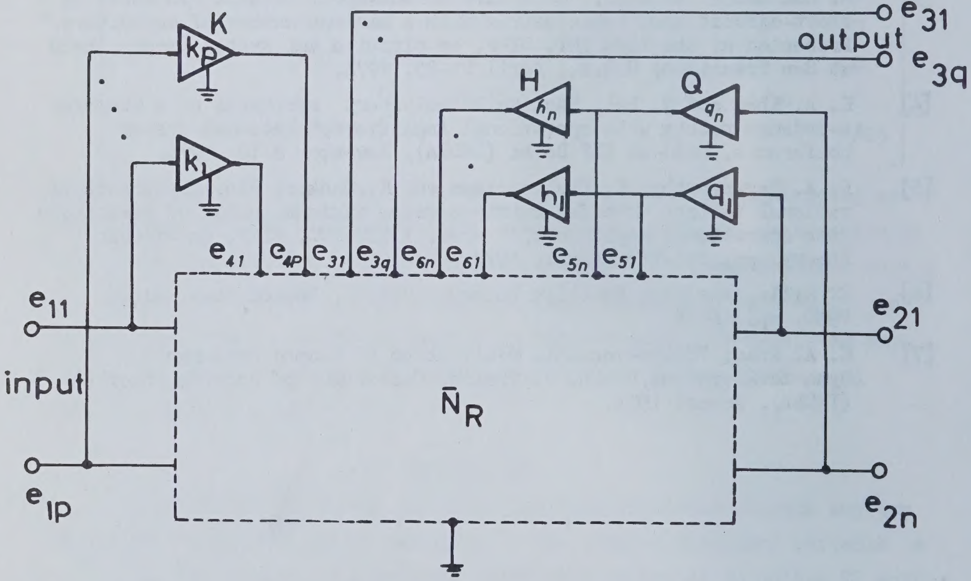
[6] E. Hille, Analytic Function Theory. Vol. I, Boston Mass. Ginn. 1959, pp. 46-58.

[7] K. A. Khan, "State-variable realization of Lumped networks & Dynamical systems," Ph. D. Thesis, University of Roorkee, Roorkee (INDIA), August 1974.



N_R
N_C
$+ e_{11}$
$+ e_{1p}$
$+ e_{21}$
$+ e_{2n}$
$-$
$-$
$-$
$-$

FIG. 1(a), NETWORK BLOCK DIAGRAM.



K
k_p
k_1
H
h_n
h_1
Q
q_n
q_1
e_{31}
output
e_{3q}
e_{41}
e_{4p}
e_{31}
e_{3q}
e_{6n}
e_{61}
e_{5n}
e_{51}
e_{11}
input
e_{1p}
$\bar{N}_R$
e_{21}
e_{2n}

FIG. 1(b), BLOCK DIAGRAM OF N_R.

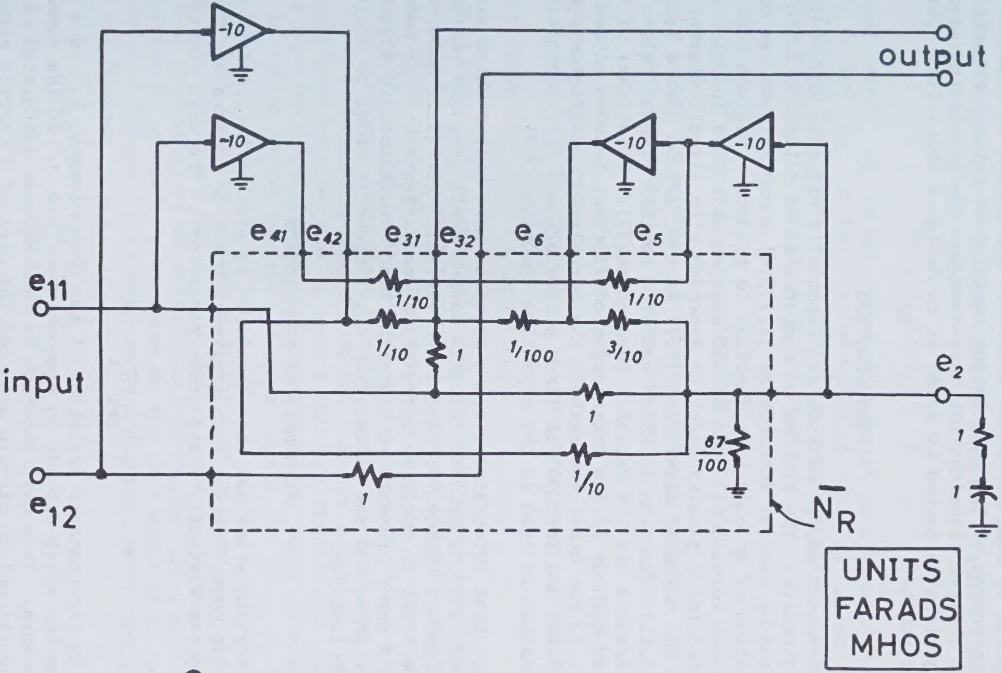

-10
-10
-10
-10
output
e41
e42
e31
e32
e6
e5
e11
input
e12
1/10
1/10
1/10
1
1/100
3/10
1/10
1
87/100
1/10
1
e2
1
1
N̄R
UNITS
FARADS
MHOS

FIG. 2. Example : Realization of T(s).

COMPUTER-AIDED TOLERANCE AND CORRELATION COEFFICIENT ASSIGNMENT: A FIRST ATTEMPT*

ARTHUR R. THORBJORNSEN AND M. MALEK DAABOUL
Department of Electrical Engineering
The University of Toledo
Toledo, Ohio 43606

ABSTRACT

The problem of computer-aided tolerance assignment is extended to include the assignment of the correlation coefficient between two circuit elements. A general algorithm for tolerance/correlation design is presented. The algorithm is tested by using it to design a thick-film hybrid integrated circuit.

INTRODUCTION

Computer-aided design methods are especially valuable when designing integrated circuits. The problem of computer-aided tolerance assignment has been addressed by several researchers [1,2,3,4] whose methods are based on the minimization of an objective function to produce a minimum cost set of circuit element tolerances. One method uses an objective function based on the Fiacco-McCormick penalty function approach with circuit element tolerances being the variable parameters [1]. Another method uses a least (P)th type of objective function in which both the nominal element values and the element tolerances are the variable parameters [2]. In neither of the last two mentioned methods is the correlation coefficient between element values considered. In the other two methods the correlation coefficients between circuit elements are included as fixed parameters and the propagation of variance equation is used in the objective function [3,4].

The fact that integrated circuit elements have values which are correlated has been used to produce circuits having small response variabilities while the elements themselves have large variabilities [5]. However, no attempt to attempt to optimize correlation coefficients has yet been reported. This paper presents an algorithm that simultaneously adjusts element tolerances and one correlation coefficient in order to minimize a circuit cost function.

MATHEMATICAL FORMULATION

The algorithm which has been developed is similar to previous ones in that a circuit cost function is minimized subject to constraints on the circuit response variability and other parameters. The cost function used is

$$C_T = \sum_{i=1}^{n} \frac{k_i}{\sigma_i} \quad , \tag{1}$$

where σ_i is the standard deviation of the (i)th element, k_i is a proportionality factor which depends on element type, and n is the number of circuit elements. This cost function is used because integrated circuit cost is proportional to circuit area and the area of a circuit element is inversely proportional to the tolerance or standard deviation.

C_T is minimized with the parameter constraints being handled by the Fiacco-McCormick penalty function method [6]. The general objective func-

---

* This work was supported by NSF Grant ENG-74-10024.

tion thus is

$$\varepsilon = W_c \sum_{i=1}^{n} \frac{k_i}{\sigma_i} + R \sum_{j=1}^{m} \frac{W_j}{G_j} , \tag{2}$$

where G_j is the (j)th constraint equation, R is the variable penalty function coefficient, W_j is a constant weighting factor, and m is the number of constraint equations. ε is minimized repeatedly, using a gradient type optimization method, for successively smaller values of R until a constrained minimum point is reached.

In order to achieve a desired circuit response standard deviation a constraint equation of the form

$$G_1 = \sigma_{AD} - \sigma_A \tag{3}$$

may be used, where σ_A is the computed standard deviation of the circuit response, A, and σ_{AD} is the desired value of σ_A. σ_A is related to the element standard deviations and correlation coefficients by the propagation of variance equation [7] which is of the form

$$\sigma_A^2 = \sum_{i=1}^{n} \sigma_i^2 \left(\frac{\partial A}{\partial p_i}\right)^2 + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(\frac{\partial A}{\partial p_i}\right)\left(\frac{\partial A}{\partial p_j}\right) \sigma_i \sigma_j r_{ij} , \tag{4}$$

where r_{ij} is the correlation coefficient between elements i and j. Equation (4) may be written in matrix form as

$$\sigma_A^2 = \underset{\sim}{x}^T \underset{\sim}{C} \underset{\sim}{x} , \tag{5}$$

where $\underset{\sim}{x}^T = \left(\frac{\partial A}{\partial p_1}, \frac{\partial A}{\partial p_2}, \cdots, \frac{\partial A}{\partial p_n}\right)$ and $\underset{\sim}{C}$ is the covariance matrix for the set of circuit elements, $\underset{\sim}{p}^T = (p_1, p_2, \cdots, p_n)$.

In this method the element standard deviations and correlation coefficients are to be adjusted to produce a desired response standard deviation. Since equation (4) is a quadratic form it is necessary that the element correlation coefficient matrix remain positive definite during the adjustment procedure. Therefore, several constraint equations must be used to keep the correlation coefficient matrix positive definite. Other constraint equations must be used to ensure that the correlation coefficient matrix is physically realizable. If necessary, the range of element standard deviation values may also be constrained.

Research is currently underway on how to form constraint equations that will keep the correlation coefficient matrix positive definite and also on the physical realizability of correlation coefficient matrices. Therefore, this algorithm is presently restricted to the case where there is only one correlation coefficient. With only one correlation coefficient present, equation (4) becomes

$$\sigma_A^2 = \sum_{i=1}^{n} \sigma_i^2 \left(\frac{\partial A}{\partial p_i}\right)^2 + 2\left(\frac{\partial A}{\partial p_e}\right)\left(\frac{\partial A}{\partial p_f}\right) \sigma_e \sigma_f r_{ef} . \tag{6}$$

In previous experimental work it was found that the correlation coefficient between thick-film resistors is always positive and never exceeds a value of about 0.95 [8]. Restricting r_{ef} to be greater than zero and less than 0.95 ensures that equation (6) will remain positive.

The complete objective function for the case of one correlation coefficient is

$$\varepsilon = W_c \sum_{i=1}^{n} \frac{k_i}{\sigma_i} + \frac{RW_1}{(\sigma_{AD}-\sigma_A)} + RW_2\left(\frac{1}{r_{ef}} + \frac{1}{r_u-r_{ef}}\right) , \tag{7}$$

where r_u is the upper limit on the correlation coefficient and is equal to 0.95. The gradient vector for the objective function is

$$\underset{\sim}{\nabla\varepsilon}^T = \left[\frac{\partial\varepsilon}{\partial\sigma_1}, \frac{\partial\varepsilon}{\partial\sigma_2}, \ldots, \frac{\partial\varepsilon}{\partial\sigma_n}, \frac{\partial\varepsilon}{\partial r_{ef}}\right] . \tag{8}$$

The partial derivatives of ε with respect to the standard derivations of uncorrelated elements are

$$\frac{\partial\varepsilon}{\partial\sigma_i} = \frac{-W_c}{\sigma_i^2} + \frac{RW_1}{(\sigma_{AD}-\sigma_A)^2}\left(\frac{\partial\sigma_A}{\partial\sigma_i}\right) , \tag{9}$$

where

$$\frac{\partial\sigma_A}{\partial\sigma_i} = \left(\frac{\sigma_i}{\sigma_A}\right)\left(\frac{\partial A}{\partial p_i}\right)^2 . \tag{10}$$

The partial derivatives of ε with respect to the standard deviations of correlated elements are of the same form as equation (9) but with

$$\frac{\partial\sigma_A}{\partial\sigma_i} = \left(\frac{\sigma_i}{\sigma_A}\right)\left(\frac{\partial A}{\partial p_i}\right)^2 + \left(\frac{2\sigma_j}{\sigma_A}\right)(r_{ij})\left(\frac{\partial A}{\partial p_i}\right)\left(\frac{\partial A}{\partial p_j}\right) . \tag{11}$$

The partial derivative of the objective function with respect to the correlation coefficient is

$$\frac{\partial\varepsilon}{\partial r_{ij}} = \frac{KW_1}{(\sigma_{AD}-\sigma_A)^2}\left(\frac{\partial\sigma_A}{\partial r_{ij}}\right) + KW_2\left[\frac{-1}{r_{ij}^2} + \frac{1}{(r_u-r_{ij})^2}\right] , \tag{12}$$

where

$$\frac{\partial A}{\partial r_{ij}} = \frac{\sigma_i\sigma_j}{\sigma_A}\left(\frac{\partial A}{\partial p_i}\right)\left(\frac{\partial A}{\partial p_j}\right) . \tag{13}$$

THE DESIGN ALGORITHM

The tolerance/correlation design algorithm proceeds as follows:

(a) Given a circuit design with nominal element values, choose a set of element standard deviations and a correlation coefficient such that the computed circuit response standard deviation, σ_A, is less than the desired value, σ_{AD}. Select a starting value for R and values for W_c, W_1, and W_2.

(b) Minimize the objective function, equation (7), for the current value of R using the current starting set of variable parameters, $\underset{\sim}{\sigma}_{po}$ and r_{efo}. The final set of parameter values produced by this minimization will be the starting set for the next iteration.

(c) Decrease R by a factor of 5.0 and repeat step (b).

(d) Repeat steps (b) and (c) until the change in the minimum ε from one iteration to the next is less than a specified $\Delta\varepsilon$, or until some other convergence criteria is met.

(e) Perform a Monte Carlo analysis using the final set of parameter values from step (d). If the final response variability, or standard deviation, is not satisfactory, it will be necessary to adjust the values of σ_{AD}, W_c, W_1, and W_2 and then start over at step (b).

Steps (b), (c), and (d) of the algorithm are performed by a single computer program. The Monte Carlo analysis of step (e) is performed by a separate program, although all four steps could be combined in a single program as in a previous method [3].

EXAMPLE

The above algorithm was tested by using it to design a thick-film hybrid integrated circuit. The circuit chosen for this test is the shunt-series feedback-pair current amplifier [9] shown in Figure 1. This circuit is especially suitable because only one correlation coefficient need be considered since the current gain is approximately equal to the ratio of two resistors,

$$A_I \simeq \frac{R_f}{R_e} .$$

Prior to using the algorithm, a shunt-series amplifier circuit was fabricated according to standard design rules [10] so that the results of the design using the algorithm could be better evaluated. The amplifier was designed to have a current gain of about 20 while keeping the range of resistor values small enough so that only one resistor printing and firing operation would be necessary. For each resistor the smallest dimension was set at 50 mils. The resulting resistor dimensional data is listed in Table 1 and the circuit layout is shown in Figure 2, which is drawn at a scale of 5 to 1.

A sample of 28 identical circuits was fabricated and the resistor values and current gain values measured. The variabilities (defined as three times the normalized standard deviation) and correlation coefficients of the parameters were computed and are listed in Table 2. The ratio R_f/R_e is used in place of measured current gain. This substitution was necessary because in the following experiment the yield was very small due to faulty transistor chips. Using the resistor ratio allowed a larger sample size. Referring to Table 2 we see that the variability of R_f/R_e is about 30%, which is less than half of the variability of either R_f or R_e, and is due to the high degree of correlation between R_e and R_f.

The shunt-series feedback-pair amplifier was next redesigned using the tolerance/correlation design algorithm. A goal of the circuit redesign was to produce a response variability of ±15%. The tolerance/correlation design program was run using a value of σ_{AD} that corresponds to a ±15% variability and a value of 0.95 for the upper correlation coefficient limit. The results of the computer run are listed in Table 3. The required variabilities for the remaining resistors are $\Delta RC_1 = \pm 83.6\%$ and $\Delta RC_2 \geq \pm 100\%$.

In order to fabricate resistors having the desired variabilities, and with R_e and R_f having the required correlation coefficient, use was made of previous experimental results and experiments [8]. The resulting dimensional data is listed in Table 4 and the layout of the redesigned circuit is shown in Figure 3.

A sample of 37 redesigned circuits were fabricated and parameters measured, with the computed statistics listed in Table 5. The redesigned circuit fell short of the goal of ±15% gain variability, however, the actual variability of ±18.8% is a dramatic improvement over the original variability of ±30.3%. The reason that a ±15% gain variability was not achieved is that the correlation coefficient between R_e and R_f did not reach the desired value.

CONCLUSIONS

A method has been presented for performing computer-aid tolerance and correlation coefficient assignment. Pending further research and experimental investigations the method is presently limited to the case where one correlation coefficient predominates. This paper demonstrates that the correlation coefficient can be as useful a design parameter as the circuit element nominal values and tolerances. It is also evident that further development of this method will produce a systematic procedure for optimizing all of the correlation coefficients and tolerances of circuit elements.

REFERENCES

[1] J. F. Pinel and D. A. Roberts, "Tolerance Assignment in Linear Networks using Nonlinear Programming," Proc. 1972 IEEE Int. Symp. on Circuit Theory, pp. 179-183, April 1972.

[2] J. W. Bandler and P. C. Liu, "Automated Network Design with Optimal Tolerances," Proc. 1973 IEEE Int. Symp. on Circuit Theory, pp. 181-184, April 1973.

[3] A. R. Thorbjornsen and S. W. Director, "Computer-Aided Tolerance Assignment for Linear Circuits with Correlated Elements," IEEE Trans. on Circuit Theory, vol. CT-20, no. 5, pp. 518-524, September 1973.

[4] S. T. Li, J. L. Hammond, and K. L. Su, "Computer-Aided Tolerance Assignment and Analysis," Research Report No. E21-620-LI-1, Georgia Institute of Technology, Chapter V, April 1974.

[5] H. R. Camenzind and A. B. Grebene, "An Outline of Design Techniques for Linear Integrated Circuits," IEEE J. of Solid-State Circuits, vol. SC-4, pp. 110-122, June 1969.

[6] A. V. Fiacco and G. P. McCormick, Nonlinear Programming: Sequential Unconstrained Minimization Techniques. New York: Wiley, 1968.

[7] D. G. Mark and L. H. Stember, Jr., "Variability Analysis," Electro-Technology, July 1965, pp. 37-48.

[8] A. R. Thorbjornsen, "Computer Aided Tolerance Assignment for Microelectronic Circuits," Technical Report No. 7612, The University of Toledo, Department of Electrical Engineering, June 1976.

[9] M. S. Ghausi, Electronic Circuits. New York: Van Nostrand Reinhold, 1971, p. 349.

[10] C. A. Harper, Handbook of Thick-Film Hybrid Microelectronics. New York: McGraw-Hill, 1974, Chapter 1.

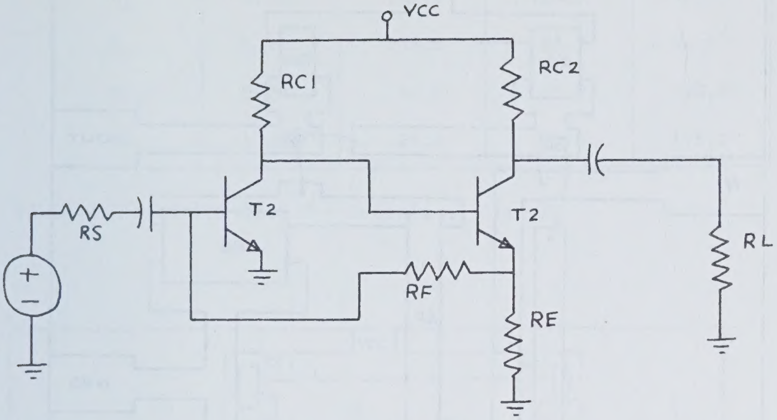


Figure 1. Shunt-Series Feedback-Pair Amplifier

Resistor	Aspect Ratio	Length (mils)	Width (mils)	Area $(\text{mils})^2$
R_e	0.291	50.0	172.0	8600.
Rf	5.70	285.0	50.0	14250.
RC_1	0.70	50.0	71.4	3570.
RC_2	1.03	51.5	50.0	2575.

Table 1. Resistor Data for Standard Design

Parameter	Variability	Correlation Coefficient
RC_1	±59.6%	- -
RC_2	±62.6%	- -
R_e	±61.5%	0.877
R_f	±68.7%	
R_f/R_e	±30.3%	

Table 2. Variabilities and Correlation Coefficient

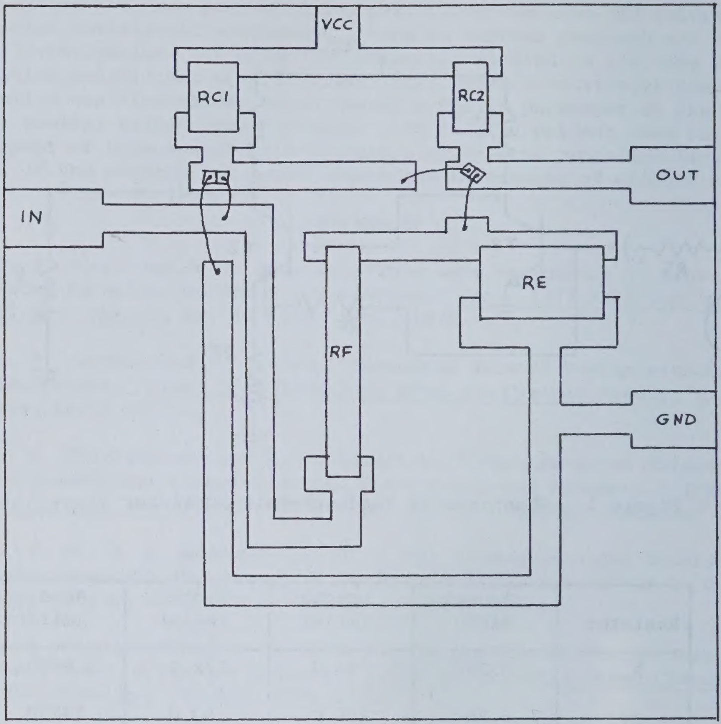


Figure 2. Circuit Layout Using Standard Design Rules

Iteration	Resistor Variability		Penalty Function Coefficient	Correlation Coefficient	Response Variability
	R_e	R_f			
1	±9.02%	±9.49%	1.000	0.888	±4.43%
2	16.5	18.0	0.200	0.916	7.14
3	26.0	28.1	0.040	0.933	10.1
4	32.9	35.7	0.008	0.942	12.1
5	38.3	41.4	0.0016	0.946	13.5
6	41.1	44.5	3.2×10^{-4}	0.948	14.3
7	42.7	46.1	6.4×10^{-5}	0.949	14.7
8	43.4	46.9	1.28×10^{-5}	0.950	14.9
9	43.7	47.2	2.56×10^{-6}	0.950	14.9
10	43.9	47.3	5.12×10^{-7}	0.950	15.0

Table 3. Results of Tolerance/Correlation Design Algorithm

Resistor	Aspect Ratio	Length (mils)	Width (mils)	Area (mils)^2
R_e	0.291	120.6	414.5	$5x10^4$
R_f	5.70	533.9	93.7	$5x10^4$
RC_1	0.70	21.0	30.0	138.4
RC_2	1.30	20.6	20.0	115.2

Table 4. Resistor Data for Redesigned Circuit

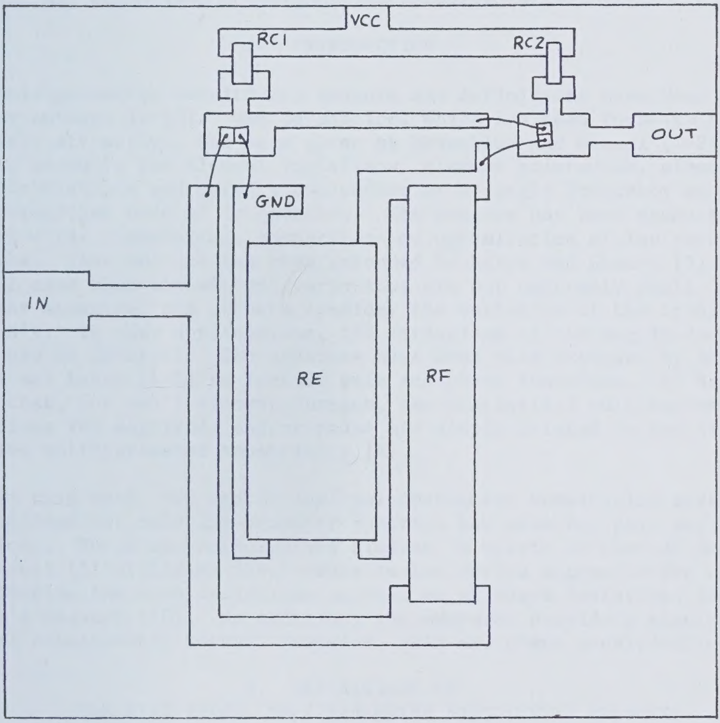

VCC
RC1
RC2
OUT
GND
IN
RE
RF

Figure 3. Layout for Redesigned Circuit

Parameter	Variability	Correlation Coefficient
RC_1	±87.45%	- -
RC_2	±107.54%	- -
R_e	±61.24%	0.911
R_f	±58.62%	
R_f/R_e	±18.8%	- -

Table 5. Variabilities and Correlation Coefficient for Redesigned Circuit

STATISTICAL MULTIPARAMETER SENSITIVITY MEASURES FOR TRANSFER, GAIN AND PHASE FUNCTIONS

CEVDET ACAR*
Department of Electrical and Computer Engineering
Wayne State University, Detroit, Michigan 48202

ABSTRACT

New statistical multiparameter sensitivity measures are defined for the transfer, gain and phase functions. The integrands of these sensitivity measures includes mean values as well as variances of element variations. It is shown that a very simple relationship exists between transfer, gain and phase measures even though element variations are large. General formulas are derived for small element variations, which would simplify evaluation of the measures. These formulas are helpful in optimization and comparison of various low sensitivity networks.

1. INTRODUCTION

Multiparameter sensitivity measure and definitions have been proposed by many authors [1-15]. One definition, which has been found to be particularly attractive, has been given by Rosenblum and Ghausi [1-2]. This measure accounts for element variations, element tolerances, element tolerance correlations and permits evaluation at a single frequency as well as at a prescribed band of frequencies. The measure has been shown to be effective for computation, comparison and optimization of low sensitivity networks. This measure has been extended by Laker and Ghausi [3] for a general case when the element variations are not extremely small. In both of these measures, the authors consider the variation of the transfer function only. In many applications, the variations of the magnitude and/or phase are of interest. The measures have been also extended by Acar, Ghausi and Laker [4-5] to include gain and phase functions. It has been shown that, for small element changes, the statistical multiparameter variations for magnitude and/or phase are simply related to the transfer function multiparameter sensitivity [4].

In this work, new statistical multiparameter sensitivity measures have been defined not only for transfer function but also for gain and phase functions. These new measures are similar in spirit to that of Rosenblum and Ghausi [1] utilizing first-order Taylor series approximation and characterize the mean variations as well as standard deviations like Shenoi's measure [10]. In addition, the measures provide a simple and general relationship between transfer, gain and phase sensitivities.

2. DEFINITIONS OF
THE STATISTICAL MULTIPARAMETER SENSITIVITY MEASURES

Let the transfer function of a given circuit be represented as

$$T = T(j\omega, x) \tag{1}$$

where x denotes the parameter vector of the circuit. Let the logarithmic variation in the transfer function due to variation in the parameter vector be represented as

---

*C. Acar is on sabbatical leave from the Technical University of Istanbul, Istanbul, Turkey.

$$\Delta\gamma = \gamma(j\omega,\ x+\Delta x) - \gamma(j\omega,\ x) \tag{2}$$

where

$$\gamma = \gamma(j\omega,\ x) = \ell n T(j\omega,\ x) \tag{3}$$

and Δx denotes the variation in x. Using the above notations, the statistical multiparameter sensitivity measure of transfer function is defined as

$$M_\gamma = \int_{\omega_1}^{\omega_2} (k_1\sigma_\gamma^2 + k_2\mu_\gamma\mu_\gamma^*)d\omega \tag{4}$$

where $\omega\varepsilon[\omega_1,\ \omega_2]$ define a designer specified frequency band. k_1 and k_2 are weighting factors. "*" denotes a complex conjugate. μ_γ and σ_γ^2 are respectively mean value and variance of $\Delta\gamma$:

$$\mu_\gamma = E[\Delta\gamma] \tag{5}$$

$$\sigma_\gamma^2 = E[\Delta\gamma\Delta\gamma^*] - E[\Delta\gamma]E[\Delta\gamma^*] \tag{6}$$

The statistical sensitivity measures of gain and phase functions can be defined in similar form. Let the gain and phase functions be represented as

$$\alpha = \alpha(\omega,\ x) = \ell n\ |T(j\omega,\ x)|, \tag{7a}$$

$$\beta = \beta(\omega,\ x) = \arg\ T(j\omega,\ x). \tag{7b}$$

Let the variations in gain and phase functions due to variation in the parameter vector be represented as

$$\Delta\alpha = \alpha(\omega,\ x+\Delta x) - \alpha(\omega,\ x), \tag{8a}$$

$$\Delta\beta = \beta(\omega,\ x+\Delta x) - \beta(\omega,\ x). \tag{8b}$$

Using these notations the sensitivity measures of gain and phase functions are defined as

$$M_\alpha = \int_{\omega_1}^{\omega_2} (k_1\sigma_\alpha^2 + k_2\mu_\alpha^2)d\omega \tag{9a}$$

$$M_\beta = \int_{\omega_1}^{\omega_2} (k_1\sigma_\beta^2 + k_2\mu_\beta^2)d\omega \tag{9b}$$

where μ_α (or μ_β) and σ_α^2 (or σ_β^2) are respectively mean value and variance of $\Delta\alpha$ (or $\Delta\beta$):

$$\mu_\alpha = E\ [\Delta\alpha], \qquad \mu_\beta = E\ [\Delta\beta] \tag{10}$$

$$\sigma_\alpha^2 = E[(\Delta\alpha)^2] - E^2[\Delta\alpha], \qquad \sigma_\beta^2 = E[(\Delta\beta)^2] - E^2[\Delta\beta] \tag{11}$$

There exists the following relation between variations in γ, α and β:

$$\Delta\gamma = \Delta\alpha + j\Delta\beta \tag{12}$$

Therefore, the mean value and variance of $\Delta\gamma$ can be written in terms of mean values and variances of $\Delta\alpha$ and $\Delta\beta$ as follows:

$$\mu_\gamma = \mu_\alpha + j\mu_\beta \text{ ,} \tag{13}$$

$$\sigma_\gamma^2 = \sigma_\alpha^2 + \sigma_\beta^2 \text{ .} \tag{14}$$

These relations between mean values and variances implie that*

$$M_\gamma = M_\alpha + M_\beta \text{ .} \tag{15}$$

It means that no matter how large element variations are, for the same specified frequency band and weighting factors, a very simple relationship exists between the sensitivity measures. The sensitivity measure of transfer function can be calculated by the use of eq. (15) once the sensitivity measures of gain and phase functions are calculated. Hence, in the following only the evaluation procedures for gain and phase measures are presented.

3. EVALUATION OF THE GAIN AND PHASE MEASURES

In order to evaluate the statistical sensitivity measures of gain and phase functions, first the variation in the network transfer function is considered. Expanding $\Delta\gamma$ into a Taylor series about $\hat{\Delta}x$ and truncating after the linear term yield

$$\Delta\gamma = S_\gamma^t \, \hat{\Delta}x \text{ .} \tag{16}$$

Here $\hat{\Delta}x$ is a column vector denoting normalized element changes in the network:

$$\hat{\Delta}x = \left[\frac{\Delta x_1}{x_1} \; \frac{\Delta x_2}{x_2} \cdots \frac{\Delta x_k}{x_k}\right]^t \tag{17}$$

where k shows the number of the elements in the network, and t transpose. Moreover S_γ is the first order multiparameter sensitivity vector of transfer function:

$$S_\gamma = \left[\frac{\partial \ell nT}{\partial \ell nx_1} \; \frac{\partial \ell nT}{\partial \ell nx_2} \cdots \frac{\partial \ell nT}{\partial \ell nx_k}\right]^t \text{ .} \tag{18}$$

Using eq. (12), the logarithmic variation in the transfer function can be expressed as

$$\Delta\gamma = S_\alpha^t \, \hat{\Delta}x + j \, S_\beta^t \, \hat{\Delta}x \tag{19}$$

---

*It is noted that, when the element variations are small such that a linear Taylor series approximation is valid, the relation between sensitivity measures defined by Rosenblum-Ghausi and Acar-Ghausi is the same as eq. (15). However, for large element change, if Laker-Ghausi definition is used, the simple relation between measures no longer exists.

where S_α and S_β are respectively real and imaginary parts of S_γ:

$$S_\gamma = S_\alpha + j\ S_\beta \tag{20}$$

In other words, S_α and S_β are respectively sensitivity vectors of gain and phase functions, namely:

$$S_\alpha = [\frac{\partial\alpha}{\partial \ell n x_1} \quad \frac{\partial\alpha}{\partial \ell n x_2} \cdots \frac{\partial\alpha}{\partial \ell n x_k}]^t \tag{21a}$$

$$S_\beta = [\frac{\partial\beta}{\partial \ell n x_1} \quad \frac{\partial\beta}{\partial \ell n x_2} \cdots \frac{\partial\beta}{\partial \ell n x_k}]^t \tag{21b}$$

Considering eq. (19), gain and phase variations can be written as

$$\Delta\alpha = S_\alpha^t\ \hat{\Delta}x \tag{22a}$$

$$\Delta\beta = S_\beta^t\ \hat{\Delta}x \tag{22b}$$

Using eqns. (10), (11) and (22) mean values and variances of $\Delta\alpha$ and $\Delta\beta$ are obtained:

$$\mu_\alpha = S_\alpha^t\ \mu \tag{23a}$$

$$\mu_\beta = S_\beta^t\ \mu \tag{23b}$$

$$\sigma_\alpha^2 = S_\alpha^t P S_\alpha \tag{24a}$$

$$\sigma_\beta^2 = S_\beta^t P S_\beta \tag{24b}$$

Here μ and P are respectively mean value vector and covariance matrix of $\hat{\Delta}x$:

$$\mu = [\ \mu_1\ \mu_2 \cdots \mu_k\ ]^t \tag{25}$$

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{bmatrix} \tag{26}$$

where

$$\mu_i = E\ [\frac{\Delta x_i}{x_i}] \qquad i = 1, 2, 3, \ldots, k \tag{27}$$

$$P_{ij} = \begin{cases} E\left[\left(\frac{\Delta x_i}{x_i}\right)^2\right] - E^2\left[\frac{\Delta x_i}{x_i}\right] & i = j \\ E\left[\left(\frac{\Delta x_i}{x_i}\right)\left(\frac{\Delta x_j}{x_j}\right)\right] - E\left[\frac{\Delta x_i}{x_i}\right] E\left[\frac{\Delta x_j}{x_j}\right] & i \neq j \end{cases} \tag{28}$$

The μ and P are known from the element change statistics. Substitution of eqs. (23) and (24) into eqs. (9) yield the following general formulas for the statistical multiparameter sensitivity measures of the gain and phase functions:

$$M_\alpha = \int_{\omega_1}^{\omega_2} S_\alpha^t \, L \, S_\alpha d\omega \tag{29a}$$

$$M_\beta = \int_{\omega_1}^{\omega_2} S_\beta^t \, L \, S_\beta d\omega \tag{29b}$$

where L is defined as

$$L = k_1 P + k_2 \, \mu \, \mu^t \tag{30}$$

These formulas are very convenient and suitable for computer-aided evaluation. For calculation of M_α and M_β one needs the matrix L and the vectors S_α and S_β. The matrix L can be directly calculated from eq. (30) since the mean value vector μ and the covariance matrix P are known from element change statistics. The sensitivity vectors S_α and S_β can be calculated by the following formulas

$$S_{\alpha i} = C_i^t \, F_\alpha \tag{31a}$$

$$S_{\beta i} = C_i^t \, F_\beta \tag{31b}$$

Here $S_{\alpha i}$ and $S_{\beta i}$ are respectively i^{th} component of the sensitivity vectors S_α and S_β. C_i is defined as follows

$$C_i = x_i \frac{\partial}{\partial x_i} \begin{bmatrix} b \\ a \end{bmatrix} = x_i \frac{\partial}{\partial x_i} \left[\begin{array}{c} b_m \\ \vdots \\ b_o \\ \hdashline a_{n-1} \\ \vdots \\ a_o \end{array}\right] \tag{32}$$

where b and a are the column vectors whose elements are the coefficients of the numerator polynomial N and denominator polynomial D of the transfer function T. F_α and F_β are real and imaginary parts of the vector F given by

$$F = \begin{bmatrix} B \\ -A \end{bmatrix} \tag{33}$$

where

$$B = \frac{1}{N}\,[(j\omega)^m \;\; (j\omega)^{m-1} \;\; 1]^t, \tag{34a}$$

$$A = \frac{1}{D}\,[(j\omega)^{n-1} \;\; (j\omega)^{n-2} \;\; 1]^t. \tag{34b}$$

It is possible to derive more useful formulas for gain and phase measures using eqs. (31). The statistical sensitivity measures of gain and phase functions in eqs. (29) can be written in the explicit forms as follows

$$M_\alpha = \int_{\omega_1}^{\omega_2} \Big(\sum_{i=1}^{k} \sum_{j=1}^{k} S_{\alpha i} S_{\alpha j} L_{ij} \Big) d\omega, \qquad M_\beta = \int_{\omega_1}^{\omega_2} \Big(\sum_{i=1}^{k} \sum_{j=1}^{k} S_{\beta i} S_{\beta j} L_{ij} \Big) d\omega \tag{35}$$

where

$$L_{ij} = k_1 \, p_{ij} + k_2 \, \mu_i \mu_j \tag{36}$$

The sensitivity products $S_{\alpha i} S_{\alpha j}$ and $S_{\beta i} S_{\beta j}$ in eqs. (35) can be written as:

$$S_{\alpha i}\, S_{\alpha j} = C_i^t \, E_\alpha \, C_j, \qquad S_{\beta i}\, S_{\beta j} = C_i^t \, E_\beta \, C_j \tag{37}$$

where E_1 and E_2 are square symmetric matrices given by

$$E_\alpha = F_\alpha \, F_\alpha^t, \qquad E_\beta = F_\beta \, F_\beta^t \tag{38}$$

Substitution of eqs. (37) into eqs. (35) yields the following formulas for the statistical multiparameter sensitivity measures of gain and phase functions:

$$M_\alpha = \sum_{i=1}^{k} \sum_{j=1}^{k} C_i^t \, K_\alpha \, C_j \, L_{ij}, \qquad M_\beta = \sum_{i=1}^{k} \sum_{j=1}^{k} C_i^t \, K_\beta \, C_j \, L_{ij} \tag{39}$$

where K_α and K_β are the integrals of the matrices E_α and E_β over radian frequency interval $[\omega_1, \omega_2]$ of our interest defined as

$$K_\alpha = \int_{\omega_1}^{\omega_2} E_\alpha \, d\omega \;, \qquad K_\beta = \int_{\omega_1}^{\omega_2} E_\beta \, d\omega \tag{40}$$

The matrices K_α and K_β are square and symmetric matrices and their elements are real numbers.

Note that the vectors C_i's are functions of the network topology and parameters while K_α and K_β are only functions of the transfer function. Hence for comparison of the sensitivity of various networks realizing a given transfer function, we need to calculate K_α and K_β only once and then C_i's are obtained for the individual networks.

The evaluation procedure of the sensitivity measures is briefly described as follows:

1. Evaluate the matrix $L = [L_{ij}]$ from element variation statistics for a given set of weighting factors using eq. (30).
2. Evaluate the vectors C_i, $i = 1, 2, 3, \ldots, k$ from design equations of the network using eq. (32).
3. Evaluate the vectors B, A, F, F_α and F_β from given transfer function using eqs. (34) and (33).
4. Evaluate the matrices E_α, E_β, K_α and K_β using eqs. (38) and (40).
5. Evaluate S_α and S_β using eqs. (31).
6. Evaluate M_α and M_β using eqs. (29) or (39).
7. Evaluate M_γ from M_α and M_β using eq. (15).

Note that if the formulas in eqs. (29) are used, then step 4 is unnecessary, if the formulas in eqs. (39) are used, then step 5 is unnecessary.

4. EXAMPLE

As an application of the procedures presented in this work, the sensitivity measures of the gain, phase and transfer functions for Figure 1 realizing the low-pass second order Butterworth function are evaluated. The transfer function of this network, assuming ideal finite gain amplifiers, is

$$T(s) = \frac{K_1K_2G_1S_1G_2S_2}{s^2 + [G_1S_1 + G_2\,S_2(1-K_1K_2)]s + G_1S_1G_2S_2}$$

where $G_i = 1/R_i$, $S_i = 1/C_i$, $i = 1, 2$.

The design equations for Butterworth function are

$$b_o = K_1K_2G_1S_1G_2S_2 = 1$$

$$a_1 = G_1S_1 + G_2S_2(1-K_1K_2) = \sqrt{2}$$

$$a_o = G_1S_1G_2S_2 = 1$$

$$K_1 = K_2 = 1$$

The vectors x and $y = [b^t a^t]$

$$x = [G_1\ S_1\ G_2\ S_2\ K_1\ K_2]^t, \quad y = [b_o\ a_1\ a_o]^t$$

The matrices K_α and K_β are calculated from the matrices E_α and E_β using eqs. (40) for bandpass region [0, 1]*.

---

* For those calculation Trapezoidal rule is used with step size 0.02.

$$K_\alpha = \begin{bmatrix} 1 & -0.345 & -0.623 \\ -0.345 & 0.184 & 0.133 \\ -0.623 & 0.133 & 0.495 \end{bmatrix} \qquad K_\beta = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.060 & -0.133 \\ 0 & -0.133 & 0.372 \end{bmatrix}$$

The vectors C_i, $i = 1, \ldots, 6$ are evaluated from design equations using eq. (32):

$$C_1 = C_2 = [1 \quad \sqrt{2} \quad 1]^t$$

$$C_3 = C_4 = [1 \quad 0 \quad 1]^t$$

$$C_5 = C_6 = [1 \quad -\frac{1}{\sqrt{2}} \quad 0]^t$$

The statistical sensitivity measures M_α and M_β are calculated from the formulas in eqs. (39), the statistical sensitivity measure M_γ from eq. (15). For this calculation the weighting factors are chosen as $k_1 = k_2 = 100$ the element change statistics are assumed as standard deviation $\sigma_i = 0.01$, correlation coefficient $\rho_{ij} = 0, \pm 0.7$, the mean value $\mu_i = 0.01, \pm 0.01$. The results are indicated in the Table 1. It is concluded by inspection of the Table that the best sensitivity performance, among selected data sets, is obtained when there exist correlation and mean variations with opposite sign.

TABLE 1

Parameter Variation Statistics			Sensitivity Measures		
σ_i%	ρ_{ij}	μ_i%	M_α	M_β	M_γ
1.	0.	1.	0.1372	0.0559	0.1931
1.	±0.7[a]	1.	0.1335	0.0492	0.1827
1.	0.	±1.[b]	0.1001	0.0115	0.1116
1.	±0.7[a]	±1.[b]	0.0964	0.0047	0.1011

[a] Correlation coefficients +0.7 for the elements pairs (R_i, R_j) and (C_i, C_j), -0.7 for the elements pair (R_i, C_j)

[b] Mean values for resistance variations -1.% for the others +1.%.

5. CONCLUSION

A somewhat different multiparameter statistical sensitivity measures for the transfer, gain and phase functions are given than in [4,5]. The formulas and the procedures are presented for small element changes, which simplify the evaluation of the measures and comparison of the networks on the basis of sensitivity. These measures account for all element variation statistics, i.e., the variances, covariances and mean values and they are also general in the sense that they can be reduced to Rosenblum-Ghausi [2] or Acar-Ghausi [4] definitions by properly selecting the weighting factors. With slight modification, large element variations can be included in the formulation. The sum of the gain and phase measures is equal to the transfer measure even though the element variations are large.

The new measures can be used in design for optimization purposes and optimization procedure described in [1] may be applied directly for them.

6. ACKNOWLEDGEMENT

I would like to thank Professor M. S. Ghausi for introducing me to the subject of sensitivity and the National Science Foundation for partially supporting this work under grant ENG 75-10181.

REFERENCES

1. A. L. Rosenblum and M. S. Ghausi, "Multiparameter sensitivity in active RC networks", IEEE Trans. on Circuit Theory, Vol. CT-18, pp. 592-599, November 1971.
2. A. L. Rosenblum, and M. S. Ghausi, "Sensitivity minimization in active RC networks", Journal of the Franklin Institute, Vol. 294, pp. 95-111, August 1972.
3. K. R. Laker, and M. S. Ghausi, "Large change sensitivity - a dual pair of approximate statistical sensitivity measures", Journal of the Franklin Institute, Vol. 298, pp. 395-413, December 1974.
4. C. Acar, and M. S. Ghausi, "Statistical multiparameter sensitivity measure of gain and phase functions", The Third International Symposium on Network Theory, Split, Yugoslavia, pp. 169-183, September 1975.
5. C. Acar, M. S. Ghausi, and K. R. Laker, "Statistical multiparameter sensitivity measures in high Q networks", Journal of the Franklin Institute, Vol. 300, No. 4, pp. 281-297, October 1975.
6. M. L. Blostein, "Some bounds on sensitivity in RLC networks", Proc. 1st Allerton Conf. Circuit Syst. Theory, pp. 488-501, 1963.
7. S. S. Haykin and W. J. Butter, "Multiparameter sensitivity indexes of performance for linear time-invariant networks", Proc. IEEE, Vol. 117, pp. 1239-1247, December 1970.
8. R. N. Biswas, and E. S. Kuh, "A multiparameter sensitivity measure for linear system", IEEE Trans. Circuit Theory, CT-18, pp. 582-587, November 1971.
9. D. Hilberman, "An approach to the sensitivity and statistical variability of biquadratic filters", IEEE Trans. Circuit Theory, CT-20, pp. 382-390, July 1973.
10. B. A. Shenoi, "Optimum variability design and comparative evaluation of thin-film RC active filters", IEEE Trans. on Circuits and Systems, Vol. CAS-21, pp. 263-267, March 1974.
11. J. D. Schoeffler, "The synthesis of minimum sensitivity networks", IEEE Trans. on Circuit Theory, Vol. CT-14, pp. 271-276, June 1964.
12. N. Fliege, "Statistical error and sensitivity measures", Proc. of the Summer School on Circuit Theory (SSCT), Vol. 2, pp. 212-214, 1974.
13. A. J. Goldstein, and F. F. Kuo, "Multiparameter sensitivity", IRE Trans. on Circuit Theory, Vol. CT-8, pp. 177-178, June 1961.
14. B. Liu, and M. E. Van Valkenburg, "A probabilistic approach to the parameter sensitivity problem", Proc. 5th Atterton Conf. Circuit and System Theory, pp. 701-710, 1967.
15. N. Fliege, "Estimation of statistical error and sensitivity measures by means of extended Gaussian Distributions", Proceedings IEEE International Synposium of Circuits and Systems, pp. 205-208, April 1976.

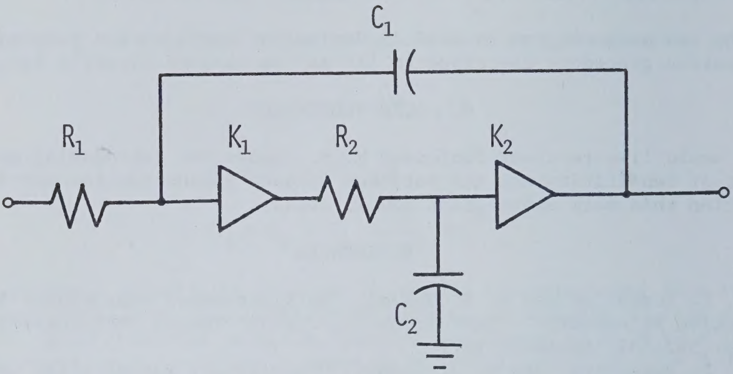

C$_1$
R$_1$
K$_1$
R$_2$
K$_2$
C$_2$

Figure 1

ON LIVENESS, DEADLOCK, AND REACHABILITY OF E-NETS

TADAO MURATA* AND TASVIR SHAH
Department of Information Engineering
University of Illinois at Chicago Circle
Chicago, Illinois 60680

ABSTRACT

Noe and Nutt modified Petri nets so as to fit their model to a real computer system, and called their graph model evaluation nets (E-nets). This paper is an intial investigation on properties of E-nets, and discusses the liveness, deadlock, and reachability of E-nets. The incidence matrix of E-nets is introduced as an analytic tool for studying the properties of E-nets.

I. INTRODUCTION

Evaluation nets or simply E-nets introduced by Noe and Nutt [1] are a modification of Petri nets. They are useful aids for simulations and performance evaluations of computer systems and programs, particularly at the task/resource level. Noe and Nutt [1] used macro E-nets for modelling a CDC 6400 computer under control of an operating system.

While E-nets have shown a powerful descriptive capability, little study has been done concerning their properties such as liveness, deadlock-freeness, reachability, equivalence, etc. This paper is an initial investigation on the properties of E-nets. In particular, the liveness, deadlock, and reachability of E-nets are discussed. Further, the incidence matrix of an E-net is introduced, which can be used as an algebraic tool for studying the behavior of E-nets.

For economy of space, it is assumed that the reader is familiar with the definitions and terminology of E-nets [1]. The additional terminology used in this paper is defined as follows. A marking M_1 in an E-net is _reachable_ from a marking M_0 if there exist a firing sequence and its associated resolution procedures that transform M_0 to M_1. An E-net is said to be _live_ for a given marking M_0 if, for _any_ marking M_1 reachable from M_0, there exist a firing sequence starting from M_1 and its associated resolution proceudures that will enable _any_ transition in the net. In order to define a directed circuit (or closed directed path) of an E-net which is analogous to a directed circuit of a digraph, transitions and (non-resolution) locations of an E-net are viewed in this paper as vertices of a digraph, rather than the E-net primitives or building blocks as was treated in[1]. A _directed_ circuit or simply dicircuit[1] of an E-net is defined as a connected subnet of the net having two or more transitions and locations in which each transition and each (inner) locations have exactly one incoming arc and exactly one outgoing arc.

---

* The author is on leave with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, 94720, for the academic year 1976-77.

II. TOKEN-INVARIANCE, LIVENESS, AND DEADLOCK

According to reference [1], there are five different transitions in an E-net: T-, F-, J-, X-, and Y-transitions. The token movements by firing T-, F- and J-transitions are the same as those in Petri nets (except that all the output locations must be empty , i.e., a token is removed from each of its input locations and is added to each of its output locations, where input and output locations of a transition are those connected by incoming and outgoing arcs of the transition, respectively. The token-count invariance of dicircuits in marked graphs [2] has played a key role in the liveness and reachability of marked graphs. So does in E-nets and we start with:

Lemma 1: The token count of a dicircuit in an E-net does not change by firing a T-, F-, or J-transition.

Proof: By the definition of E-nets, an inner location has exactly one incoming and one outgoing arcs. Thus, a transition in an E-net belongs to a dicircuit if and only if a pair of its input and output locations belongs to the dicircuit. Therefore, when a T-, F-, or J-transition belonging to a dicircuit Γ is fired, one token is removed from Γ and one token is added to Γ, resulting in zero token-count change in Γ. When a transition does not belong to a dicircuit, obviously there is no token-count change in the dicircuit.

Lemma 2: The token count of a dicircuit in an E-net remains the same or decreases by one after one firing of an X-transition.

Proof: When an X-transition is fired, a token moves from its (unique) input location to one of the two output locations depending on the resolution procedure, r_i. If an X-transition belongs to a dicircuit Γ (thus its input location also belongs to Γ) and the output location that will receive a token does not belong to Γ, then the token count of Γ decreases by one. Otherwise, it remains the same.

Lemma 3: The token-count of a dicircuit in an E-net remains the same or increases by one after one firing of a Y-transition.

Proof: The same as that of Lemma 2 except that the words "input and output locations" are interchanged, and "decrease" is replaced by "increase".

Dicircuits of an E-net can be classified into the following four types according to the existence of X-and/or Y-transitions in a dicircuit:

1) A type 1 dicircuit is one containing no X- and no Y-transitions.

2) A type 2 dicircuit is one containing X-transition(s) but no Y-transitions.

3) A type 3 dicircuit is one containing Y-transition(s) but no X-transitions.

4) A type 4 dicircuit is one containing X- and Y- transitions.

A dicircuit is said to be token-free or token-full if each of its locations has no token or has a token, respectively. The following theorem is obvious from Lemmas 1, 2, and 3:

Theorem 1: An E-net is not live (deadlocked) for a marking M_0 if there exists a token-free or token-full type 1 dicircuit, a token-free type 2 dicircuit, or a token-full type 3 dicircuit.

Note that a transition on a token-full type 1 or type 3 dicircuit cannot be fired since in an E-net a transition firing that would produce more than one token in a location is not allowed. A marking stated in Theorem 1 may be looked upon as a deadlock in the sense that it is a state in which two or more processes (executions of transition firings) are waiting indefinitely for conditions which will never hold.

By Lemma 2 (or 3), a type 2 (or type 3) dicircuit may lose (or gain) enough tokens so that there will be some transition which cannot be fired because there is no way to increase (or decrease, respectively) tokens in the dicircuit. Therefore, we have:

Theorem 2: An E-net cannot be live if there exists a type 2 or type 3 dicircuit.

The counterpositive of Theorem 2 leads to the following necessary condition for an E-net to be live.

Theorem 3: It is necessary for an E-net to be live for a given marking M_0 that each dicircuit (if any) must be of type 1 or type 4, where a type 1 dicircuit is not token-free nor token-full for M_0.

It should be noted that a token can be placed on (or removed from) peripheral input (output) locations of an E-net whenever necessary. Otherwise, the transitions directly connected to these locations could not be fired once they fire.

III. INCIDENCE MATRIX AND STATE EQUATION OF E-NETS

The incidence matrix $A = [a_{ij}]$ of an E-net with t transitions and p non-resolution locations is a t x p matrix, the typical entry of which a_{ij} is defined as follows: where the subscripts i and j denote transition i and non-resolution location j, respectively, and $\overline{r}_i$ denotes the complement of resolution procedure r_i, i.e., $\overline{r}_i = 0$ or 1 for $r_i = 1$ or 0, respectively:

1) For the two output locations of an X-transition i, $a_{ij} = r_i$ or $\overline{r}_i$

depending on whether location j is labelled "1" or "0" (on its incoming arc) respectively.

2) For the two input locations of an Y-transition i, $a_{ij} = -r_i$ or $-\bar{r}_i$ depending on whether location j is labelled "1" or "0" (on its outgoing arc), respectively.

3) For all other cases, $a_{ij} = 1$ or -1 if location j is an output or input location of transition i, respectively, and otherwise $a_{ij} = 0$.

Example 1: The incidence matrix of an E-net shown in Figure 1 is given by:

$$A = \begin{array}{c|cccccccc} & j_1 & j_2 & j_3 & j_4 & j_5 & j_6 & j_7 & j_8 \\ \hline i_1 & -1 & \bar{r}_1 & r_1 & 0 & 0 & 0 & 0 & 0 \\ i_2 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ i_3 & 0 & 0 & -1 & -1 & 0 & 1 & 0 & 0 \\ i_4 & 0 & 0 & 0 & 1 & 0 & -1 & 1 & 0 \\ i_5 & 0 & 0 & 0 & 0 & -\bar{r}_5 & 0 & -r_5 & 1 \\ i_6 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{array} \qquad (1)$$

Note that the information of resolution locations is included in the entries as r_i or $\bar{r}_i$, but not as columns of the incidence matrix A. (Also, resolution locations are not drawn explicity in the figure, but are assumed to be present). The incidence matrix A of E-nets defined in the above carries the complete information on the topology of an E-net and can be used or stored in a computer as a substitute of drawing an E-net. For example, by inspecting nonzero entries of each row of A in example 1 it is easy to recognize that transitions i_1, i_2, i_3, i_4, i_5, i_6 are X-, T-, J-, F-, Y- and T-transitions, respectively.

Furthermore, from the definition of the incidence matrix A and the firing rule of each transition, it is seen that when transition i is fired, a location j receives a token if $a_{ij} = 1$, gives away a token if $a_{ij} = -1$ or neither if $a_{ij} = 0$, where r_i for an X- or Y-transition[2] is set to either 1 or 0 immediately before its firing by the resolution procedure. In other words, row i of A represents the token changes in the p locations by one firing of transition i, and the following matrix equation can be written:

$$M_{k+1} = M_k + A^T U_k \quad , k = 0, 1, 2, \ldots \qquad (2)$$

where M_k is the marking (p x 1 column) vector of 0's and 1's, the jth entry of which denotes the number of tokens on location j immediatley before the kth firing; U_k is the firing (t x 1 column) vector of 0's and 1's, the ith entry of which is 1 if transition i is fired at the kth stage and is 0 otherwise; A^T is transpose[3] of the incidence matrix A. Equation (2) can be viewed as a state equation for a discrete-time system, where state vectors M_k

and control vectors U_k for all $k = 0, 1, 2, \ldots$ are constrained to be vectors of 0's and 1's, since transition firings that would result in more than one token or less than zero tokens in any location are not allowed in an E-net.

IV. APPLICATIONS OF THE INCIDENCE MATRIX TO LIVENESS AND REACHABILITY

From the discussion in the previous section, it can be seen that each entry of column j of A represents the token change in location j by one firing of the transition corresponding to the entry. Thus, the vector obtained by adding some columns of A represents the changes in token count of a subset of the locations corresponding to the columns by firing the respective transitions. Therefore, we have:

Theorem 4: In an E-net, the token count of a subset of m locations, j_i, $j_2, \ldots, j_m$, where $m \leq p$,

i) remains the same,

ii) remains the same or increases, or

iii) remains the same or decreases,

after any transition firing (with any resolution procedure) if and only if the sum of the corresponding columns $j_1, j_2, \ldots, j_m$ is a vector containing i) zeros only, ii) no negative entries, or iii) no positive entries, respectively, whether $r_i = 1$ or 0 for all resolution procedures r_i's.

The three subsets of locations stated in Theorem 4 are called the token-invariant set, token-increasing set, and token-decreasing set, respectively.

Example 2: The two subsets of locations, (j_4, j_6) and $(j_1, j_2, j_3, j_5, j_6, j_7, j_8)$ in the E-net shown in Figure 1 are token-invariant sets since the sums of columns corresponding to these sets in the incidence matrix A are zero vectors in both cases as is easily verified by using (1) (note that $r_i + \bar{r}_i = 1$ for all r_i). The incidence matrices of two E-nets shown in Figure 2 are given, respectively, by

$$\begin{array}{c} \\ i_1 \\ i_2 \\ i_3 \\ i_4 \end{array} \begin{array}{c} \begin{array}{ccccc} j_1 & j_2 & j_3 & j_4 & j_5 \end{array} \\ \left[\begin{array}{ccccc} 1 & 0 & 0 & -1 & -1 \\ -1 & r_2 & \bar{r}_2 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{array}\right] \end{array} \qquad (3)$$

and

$$\begin{array}{c} \\ i_1 \\ i_2 \\ i_3 \\ i_4 \end{array} \begin{array}{c} \begin{array}{ccccc} j_1 & j_2 & j_3 & j_4 & j_5 \end{array} \\ \left[\begin{array}{ccccc} 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ -r_3 & -\bar{r}_3 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 1 \end{array}\right] \end{array} \qquad (4)$$

For example, (j_1, j_2, j_3, j_4) and a type 2 dicircuit (j_1, j_3, j_5) in Figure 2 (a) are token-decreasing sets since the sums of these column sets, $[0\ 0\ 0\ -1]^T$ and $[0\ (\bar{r}_2-1)\ 0\ 0]^T$ contain no positive entries, respectively. (j_1, j_2, j_3, j_4) and a type 3 dicircuit (j_2, j_3, j_5) in Figure 2 (b) are token-increasing sets since the sums of these column sets $[0\ 1\ 0\ 0]^T$ and $[0\ 0\ 1-\bar{r}_3)\ 0]^T$ contain no negative entries, respectively.

The token-decreasing or increasing set, which contains a type 2 or type 3 dicircuit, respectively, provides an algebraic method for detecting the non-liveness of an E-net from its incidence matrix. The counterpart of Theorem 2 may be stated as follows:

Theorem 5: An E-net cannot be live if it has a token-decreasing or increasing set.

Token-invariant sets are closely related to type 1 and type 4 dicircuits, and a topological characterization of token-invariant sets may be stated as follows. Consider a connected subnet G of an E-net which has the following properties:

(i) For every T-, F- or J-transition in G, exactly one input and one output locations of the transition belong to G.

(ii) For every X- or Y-transition in G, all the three (non-resolution) locations (one input and two outputs for X, or two inputs and one output for Y) belong to G.

Then, the set of locations belonging to G is an token-invariant set. It is easily verified from the definition of each transition firing or from Theorem 4 that the token count of locations in G is invariant under any transition firing.

Some examples of the above subnet G are given below (where recall that by definition a location in an E-net has exactly one incoming and one outgoing arcs).

1) A type 1 dicircuit, or a directed path containing no X- and no Y-transitions, which starts at a peripheral (input) locations and ends at a peripheral (output) location of an E-net (if the addition or removal of a token in a peripheral location is not considered to be a transition firing).

2) Consider a directed path P containing both X- and Y-transitions, which may be an open path from a peripheral (input) location to a peripheral (output) location of an E-net, or a closed path, i.e., a type 4 dicircuit. Define a by-path associated with P as a directed path which branches out from P at an X-transition and joins back to P at a Y-transition or as a backward directed path which branches out from P at a Y-transition and joins back to P at an X-transition. Define a new P as the union of the

present P and all the by-paths associated with the present P. Continue the above process. If this process terminates with P which satisfies the above properties of G, then the set of locations in this subnet P is a token-invariant set (see Example 2).

The main motivation of introducing token-invariant sets is to relate them to the reachability of markings in an E-net. For this purpose, define a token-invariance matrix $B = [b_{ij}]$ of an E-net with p (non-resolution) locations and q different token-invariant sets, as a q x p matrix, where its entry b_{ij} is 1 if location j belongs to the ith token-invariant set, and is 0 otherwise.

It is obvious from Theorem 4 that

$$BA^T = 0 \, . \tag{5}$$

Theorem 6: If there exists a firing sequence leading from an initial marking M_0 to another marking M_n in an E-net, then

$$B\,(M_n - M_0) = 0 \, . \tag{6}$$

Proof: If there exists a firing sequence $\{U_0, U_1, \ldots, U_{n-1}\}$ which transforms M_0 to M_n, then from (2) we have

$$A^T\,(U_0 + U_1 + \ldots + U_{n-1}) = M_n - M_0 \tag{7}$$

It is seen from (5) and (7) that (6) holds.

It should be noted that B in (5) and (6) can be generalized by defining the rows of B as a set of all independent solutions y of a homogeneous equation, $Ay = 0$. Equation (6) with this B will provide a stronger necessary condition for the reachability than Theorem 6. In fact, it has been shown that a similar condition is necessary and sufficient for a live marking M_0 to reach another M_n in the case of marked graphs [5].

V. CONCLUSIONS

Properties of E-nets have been investigated for the first time. Theorems 1, 2 and 3 are concerned with the structural information on the liveness and deadlock, whereas Theorems 4, 5 and 6 are with an algebraic approach to the nonliveness and reachability. The algebraic approach is based on the incidence matrix A of an E-net introduced in this paper. Since the incidence matrix carries the complete topological description of an E-net, it can be conveniently stored in a computer, and may be used for many other studies of E-nets, in addition to those presented here.

E-nets can be regarded as an interpreted marked graph [4]. However, E-nets have a much more powerful descriptive capability and is much less amenable to analysis than marked graphs. The analysis of E-nets would be even more complex if it included the effects of transition-time delays and the internally controlled resolution procedures that are dependent on some markings in an E-net.

REFERENCES

[1] J. D. Noe and G. J. Nutt , "Macro E-nets for representation of parallel system", IEEE Trans. on Comp., Vol. C-22, No. 8, pp. 718-727, Aug. 1973

[2] F. G. Commoner, A. W. Holt, S. Even, and A. Pnueli, "Marked directed graphs", J. of Computer and System Sciences, Vol. 5, pp. 511-532, Oct. 1971.

[3] For example, N. Deo, Graph Theory with Applications to Engineering and Computer Science, Prentice-Hall, Englewood Cliffs, New Jersey, 1974.

[4] T. Murata, "State equation for E-net interpreted marked graphs", procs. of the 19th Midwest Symposium on Circuits and Systems, University of Wisconsin - Milwaukee, August, 1976.

[5] T. Murata and R. W. Church, "Analysis of marked graphs and Petri nets by matrix equations", Research Report No. M.D.C. 1.1.8, Department of Information Engineering, University of Illinois, November, 1975.

FOOTNOTES

1. A "self-loop" or a dicircuit consisting of one transition and one location was not considered in [1] and thus it is omitted here although it could be treated separately.
2. In reference [1], the resolution procedure r_i for Y-transition is not used in the case that a token arrives at either one of the two input locations. In the case of this non-simultaneous token arrival, it is assumed here that the r_i is set to 1 or 0 when a token arrives at the location labelled "1" or "0", respectively.
3. The reason for not defining this transposed matrix as the incidence matrix of an E-net is that the present one is analogous to the vertex-to-arc incidence matrix of a digraph.

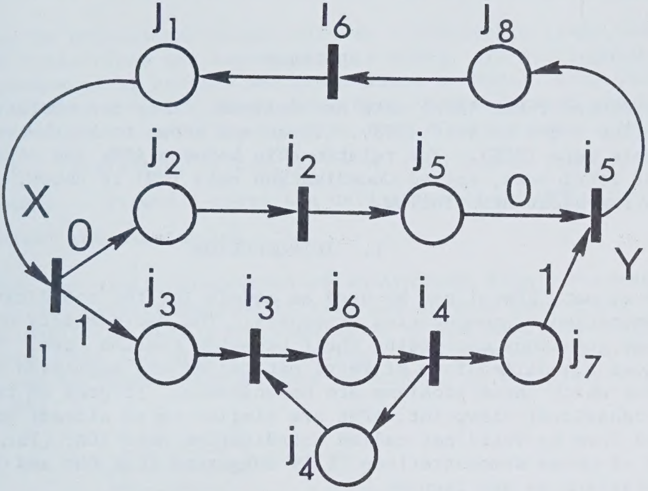
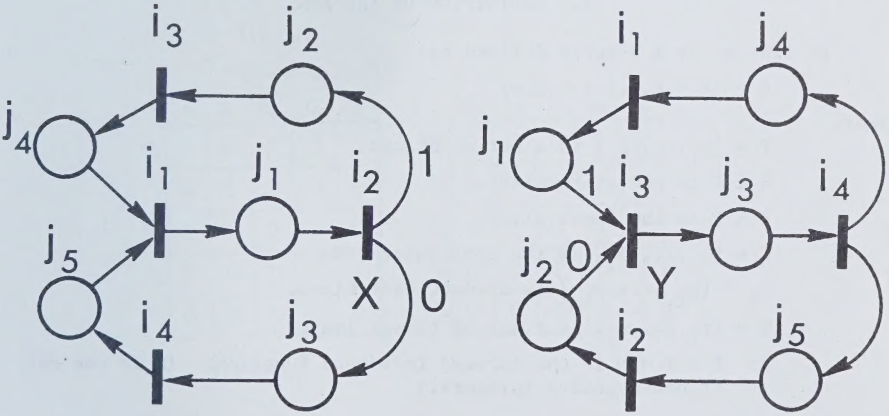

j_1
i_6
j_8
j_2
i_2
j_5
i_5
X
0
0
Y
1
j_3
i_3
j_6
i_4
i_1
1
j_7
j_4

FIGURE 1 AN EXAMPLE OF AN E-NET


i_3
j_2
j_4
i_1
j_1
i_2
1
j_5
X
0
i_4
j_3
i_1
j_4
j_1
1
i_3
j_3
i_4
j_2
0
Y
i_2
j_5

(a) (b)

FIGURE 2 ILLUSTRATION OF TOKEN-INCREASING AND DECREASING SETS

CHARACTERISTICS OF SOME AUGMENTED PETRI NETS*

J. E. SMITH† AND T. MUDGE
Coordinated Science Laboratory
University of Illinois, Urbana, Illinois 61801

ABSTRACT

Augmented Petri (APN) nets are defined. They are characterized by computation sequence sets (CSS). These are shown to be the recursively enumerable sets (RES). The relationship between APNs and other modified forms of Petri nets, called Coordination nets (CN) is shown. Several corollaries about APNs follow.

1. INTRODUCTION

Petri nets [Petr] can be used as models for the coordination and synchronization of computation processes. The decidability of several important questions concerning their behavior remains open. This paper introduces a modified form of Petri net called the augmented Petri net (APN) for which these problems are undecidable. It goes on to show that, from a behavioral viewpoint, APNs are similar to an already existing modified form by Petri net called Coordination nets (CN) [Pati]. As a by-product of these demonstrations it is suggested that CNs and Petri nets are not equivalent as was thought.

Finally, and as a consequence of their computation sequence sets (CSS) being the recursively enumerable sets (RES), the following corollaries are shown to be true for APNs:

1. The reachability problem is undecidable.
2. The boundedness problem is undecidable.
3. The coverability problem is undecidable.

2. DEFINITION OF THE APN

An APN, A, is a 9-tuple defined by:

$$A = \langle P,S,F,C,T,f,b,\Sigma,s\rangle$$

where,

$P = \{p_1,\ldots,p_k\}$ is a set of places.

$S \in P$ is the start place.

$F \in P$ is the final place.

$C = \{c_1,\ldots,c_\ell\}$ is the constraint set.

$c_i = \{p_{i_1},\ldots,p_{i_m}\}$ is a constraint class.

$T = \{t_1,\ldots,t_n\}$ is a set of transitions.

$f\colon\ P \times T \to N$ is the forward incidence function. (N is the set of non-negative integers.)

---

* This work was supported by the Joint Services Electronics Program (U.S. Army, U.S. Navy, and U.S. Air Force) under Contract DAAB-07-72-C-0259

† Currently with the Department of Electrical and Computer Engineering at the University of Wisconsin, Madison 53706.

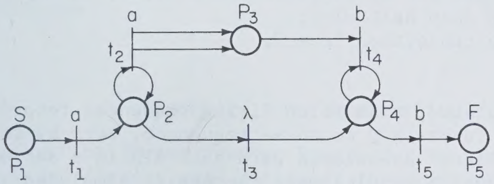
b: $P \times T \to N$ is the backward incidence function.

$\Sigma = \{s_1, \ldots, s_p\}$ is an alphabet of computations.

g: $T \to \{\Sigma \cup \lambda\}$ is the output function. (λ is the null computation.)

An APN may be represented graphically as a bi-partite graph, in which the places and transitions are represented as nodes. To distinguish them, places are represented by circles and transitions by bars. Arcs connect places to transitions and vice versa, according to the forward and backward incidence functions. If $f(p_i, t_j) = m$, then m arcs connect place p_i to transition t_j. Similarly, if $b(p_i, t_j) = n$, then n arcs connect transition t_j to place p_i. Figure 1 gives the definition of an APN together with its graphical representation.

So far just the static properties of an APN have been presented. Since an APN is an abstract machine it also has dynamic properties. These properties are exhibited during the simulation of the APN. Simulation is defined by a procedure and abstract entities called tokens, which reside at places and can be represented as dots or numbers in the place circles of the APN graph.



$P = \{P_1, P_2, P_3, P_4, P_5\}$

$S = P_1$

$F = P_5$

$C = \{P_3, P_5\}$

$T = \{t_1, t_2, t_3, t_4, t_5\}$

f:

	t_1	t_2	t_3	t_4	t_5
P_1	1	0	0	0	0
P_2	0	1	1	0	0
P_3	0	0	0	1	0
P_4	0	0	0	1	1
P_5	0	0	0	0	0

b:

	t_1	t_2	t_3	t_4	t_5
P_1	0	0	0	0	0
P_2	1	1	0	0	0
P_3	0	2	0	0	0
P_4	0	0	1	1	0
P_5	0	0	0	0	1

$\Sigma = \{a, b\}$

g:

t_1	t_2	t_3	t_4	t_5
a	a	λ	b	b

FP-5301

Figure 1. An Example APN.

Before giving the simulation procedure it is necessary to define some new terms. A transition t_i is *primed* if the number of tokens at each place p_j (written $|p_j|$) satisfies,

$$|p_j| \geq f(p_j,t_i)$$

A transition t_i is *fireable* if it is primed and the resulting new token distribution given by,

$$|p_j| \leftarrow |p_j| - f(p_j,t_i) + b(p_j,t_i) \ \forall j$$

satisfies the constraint condition viz.

$$\nexists\ c_k \in C \ (c_k = \{p_{k_1},\dots,p_{k_s}\})$$
$$\ni |p_h| > 0 \quad \text{for } h = k_1,\dots,k_s.$$

A transition that is fireable may *fire* to produce the above new token distribution or state.

The simulation procedure now follows below.

1. Initialize the APN by placing one token in the start place.
2. Compute the set of fireable transitions, U.
3. If U = $\emptyset$ then halt else:
4. Fire one transition from U.
5. Goto 2.

As an APN is simulated, transition firing sequences record the simulation. Applying the function Σ to these sequences, term by term, yields sequences of computations. Associated with each APN is a set of such sequences, any one of which results when the APN is simulated from its initial state to the state $|F| = 1$ and $p_j = 0 \ \forall j \ni p_j \neq F$. This set is called the computation sequence set [Pete] (written CSS).

Figure 1 shows an APN which has been initialized with a token in $S = p_1$. Its CSS is the sequences $a^n b^{2n-1}, n > 0$.

3. PROPERTIES OF CSSs

In this section we will show that the CSSs produced by APNs are the RES (or type 0 languages), so that APNs and Turing machines can be considered equally powerful at modelling the coordination and synchronization of computation sequences.

For this, a result of [Mati] is used. It can be stated as the following assertion:

For each S $\in$ RES, $\exists$ a diophantine polynomial (taken here to mean a polynomial with integer coefficients and integer roots) in n + 1 variables, $P(x,y_1,\dots,y_n)$, such that if x is a non-negative integer encoding of any $w \in S$ then $\exists$ integers $y_1,\dots y_n$ such that $P(x,y_1,\dots,y_n) = 0$. Furthermore $\nexists$ integers $y_1,\dots,y_n$ such that $P(x,y_1,\dots,y_n) = 0$ for any x that is not an encoding of a $w \in S$.

By showing how an APN can be made to behave like such a polynomial, it can be shown that the RES form a subset of the CSSs for APNs.

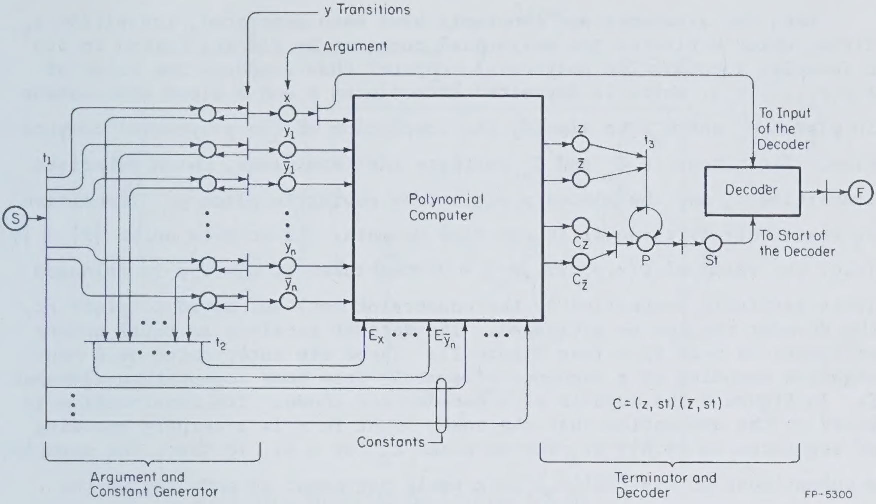

y Transitions
Argument
x
y_1
$\bar{y}_1$
y_n
$\bar{y}_n$
t_1
S
t_2
Polynomial Computer
E_x
$E_{\bar{y}_n}$
z
$\bar{z}$
t_3
C_z
$C_{\bar{z}}$
P
St
To Input of the Decoder
Decoder
F
To Start of the Decoder
C = (z, st) ($\bar{z}$, st)
Constants
Argument and Constant Generator
Terminator and Decoder
FP-5300

Figure 2. The Construction.

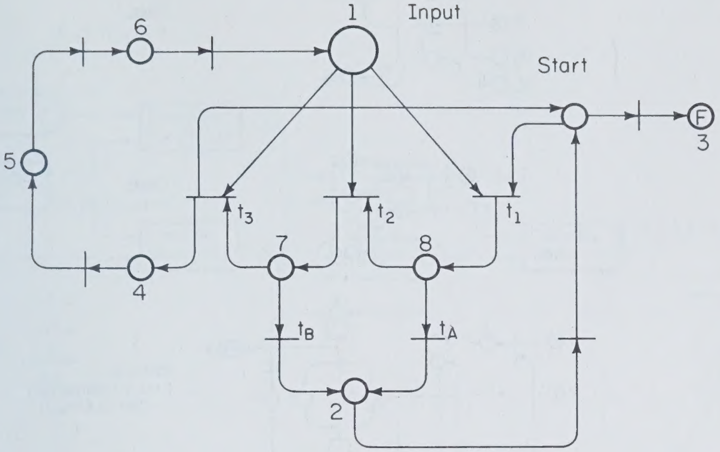
The construction is outlined in Figure 2. Initializing the start place with a token fires t_1 and activates the argument and constant generator. This causes a set of places corresponding to the integers $(x, y_1, \ldots, y_n)$ and the constants of the polynomial to be filled with tokens. So that each integer $y_i \in (y_1, \ldots, y_n)$ can range over the positive and negative integers they are encoded into two places, a positive one y_i (see Figure 2) and a negative one $\bar{y}_i$. The value of any y_i is given by the difference between the number of tokens in the positive place and the number in the negative place. Since x only ranges over the non-negative integers its value is given by the number of tokens in a single place x. The actual values of x and the y_i are determined by an arbitrary number of firings of the y transitions. This part of the machine is non-deterministic and models a search through vectors of the form $(y_1, \ldots, y_n)$ to find if there exists integer values for the y_is such that for some non-negative x, $P(x, y_1, \ldots, y_n) = 0$. Failure, for any argument $(x, y_1, \ldots, y_n)$ selected in this non-deterministic manner, results in the machine producing the λ computation sequence and not terminating properly. (Proper termination corresponds to the state $|F| = 1$ and $p_j = 0 \; \forall j \ni p_j \neq F$; see the definition of a CSS in section 2.) Success results in the machine producing a computation sequence which corresponds to an element of the RES defined by the machines polynomial. The constants produced by the argument and constant generator, for use as the integer coefficients in the polynomial computer, are encoded in a manner similar to the y_is.

Their constant value is determined by the use of multiple arcs from t_1 to their respective places.

Once the arguments and constants have been generated, transition t_2 fires, which activates the polynomial computer by placing tokens in its E (enable) inputs. The polynomial computer than computes the value of $P(x,y_1,\ldots,y_n)$, which is deposited into places z and $\overline{z}$ along with tokens in places C_z and $C_{\overline{z}}$, to signify the completion of the polynimial computation. The tokens in C_z and $C_{\overline{z}}$ initiate the terminator, which comprises transition t_3 and the places p and st, by moving to place p. Transition t_3 repeatedly fires until it can fire no more. If at this point $|z| = |\overline{z}|$ (i.e. the value of $P(x,y_1,\ldots,y_n) = 0$) the token at place p is released (this partially controlled by the constraint set) and moved to place st, the decoder can now be activated. The decoder receives an equal number of tokens as were in x (see Figure 2). These are interpreted as a non-negative encoding of a sequence of symbols from some computation alphabet Σ. In Figure 3 the details of a decoder are shown. Its construction is based on the assumption that the token count in x is a ternary encoding of sequences in $\{A,B\}^*$ (1_3 for an A and 2_3 for a B), so that, for example, a token count of 149 (12112_3) in x would represent an encoding of the sequence BAABA (reading right to left). The operation of the decoder is straightforward, the contents of place 1 are repeatedly counted down by 3, by the sequential firing of transitions t_1, t_2 and t_3. The remainder after each count down will be either one or two tokens, resulting in either t_A or t_B firing (this is partially controlled by the constraint set). The transitions t_A and t_B correspond to computations A and B, respectively, from the computation alphabet. All other transitions in the machine map to the λ computation (i.e. $g(t_A) = A$, $g(t_B) = B$ and $g(t) = \lambda \; \forall t \neq t_B, t_B$). When the sequence has been generated a token is placed in the final place F as the constraint set dictates. Of course, other encodings can be assumed and other decoders constructed in order to produce recursively enumerable sets defined on different alphabets.

Using the construction of Figure 2 any RES can in theory be generated by considering the alphabet of the set to be a computation alphabet, then selecting a non-negative encoding of that alphabet (thus defining the decoder) and then selecting the appropriate diophantine polynomial (thus defining the polynomial computer), whose integer roots coincide with those values for x that are encodings of the members of the required RES.

Elements of the polynomial computer are shown in Figure 4. There are three basic ones, an adder, a copier and a multiplier. In the adder, if x tokens are deposited in place 1, y in place 2 together with a token in E_1 and one in E_2, the adder will stop with $x + y$ tokens in place 3 and a token in place C_o. The addition will not start until a token is placed in both E_1 and E_2, and a token is not placed in C_o until places 1 and 2 are empty and the $x + y$ tokens are in place 3. The Es can be regarded as the enable places which start the basic computation and C_o as the completion place which signifies completion of the basic computation. Notice how the constraint set ensures correction operation. From the explanation of the adder's operation, the operation of the copier element should be apparent.



$C = (1,2)(1,5)(1,3)(3,4)(3,5)(3,6)(5,7)(5,8)$

FP-5299

Figure 3. A Decoder.

Operation of the multiplier is slightly more complicated. If x tokens are deposited in place 1 and y in place 2, then $x \cdot y$ are produced in place 9. The operation is synchronized by tokens in the places E_1 and E_2, and completion is signified by a token in place C_o as with adder. The element starts by moving the tokens in place 2 into place 7. These tokens are alternately moved between places 7 and 6. In transfering them from 7 to 6 transition t is fired y times causing y tokens to be deposited in place 9. This is regulated to occur exactly x times (resulting in $x \cdot y$ tokens in place 9) by moving one of the tokens in 1 to place 4 each time the tokens in 7 are to fire transition t y times. Place 8 is used to clear the y tokens from place 7 when place 1 is empty, prior to the completion of the multiplication.

These basic elements can be combined to form complex adders, copiers and multipliers (see Figure 5) so that signed arithmetic can be modelled, for those integers encoded into a positive and negative place. When combining basic elements, output places become the input places of subsequent elements. Similarly their corresponding C_o and E places are matched to preserve the overall sequencing of the computer.

Figure 6 shows how a polynomial computer for the polynomial xy + 3x - 2 can be built from complex adders, copiers and multipliers. The constants 3 and -2 would be generated by the argument and constant generator. In the case of constant 3, the firing of transition t_1 of the argument and constant generator (see Figure 2) would place 3 tokens into one of the positive input places of the complex multiplier shown at the top of

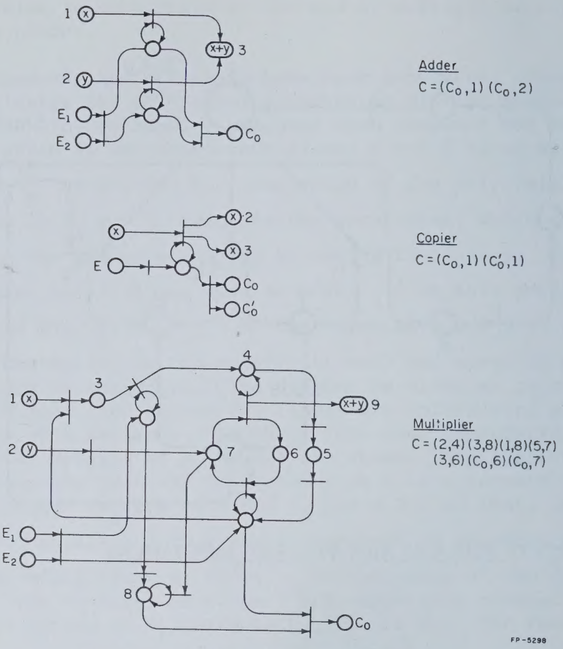
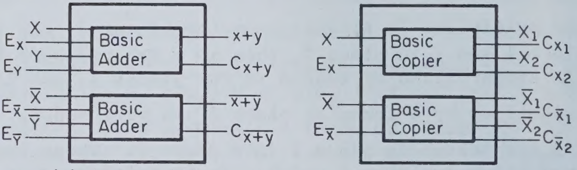

Adder
C = (C0,1) (C0,2)
Copier
C = (C0,1) (C'0,1)
Multiplier
C = (2,4) (3,8) (1,8) (5,7) (3,6) (C0,6) (C0,7)
x+y
E1
E2
C0
FP-5298

Figure 4. The Basic Elements.


Basic Adder
Basic Adder
x+y
Cx+y
Basic Copier
Basic Copier
X1
X2
Cx1
Cx2

(a) Complex Adder (b) Complex Copier

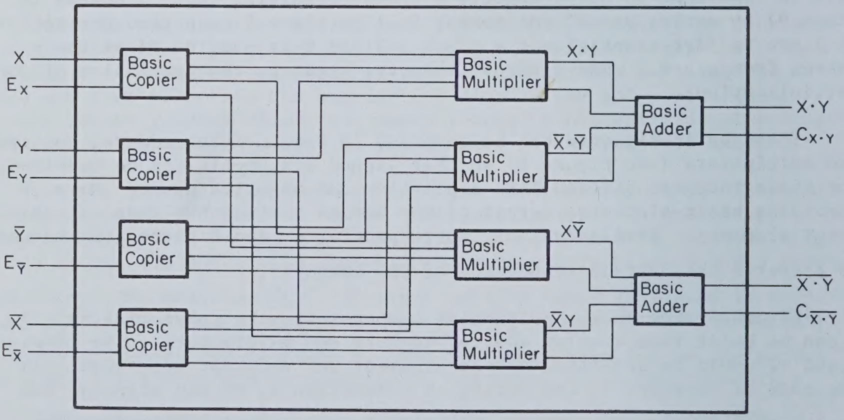

Basic Copier
Basic Multiplier
Basic Adder
X·Y
Cx·y
FP-5297

Figure 5. The Complex Elements.

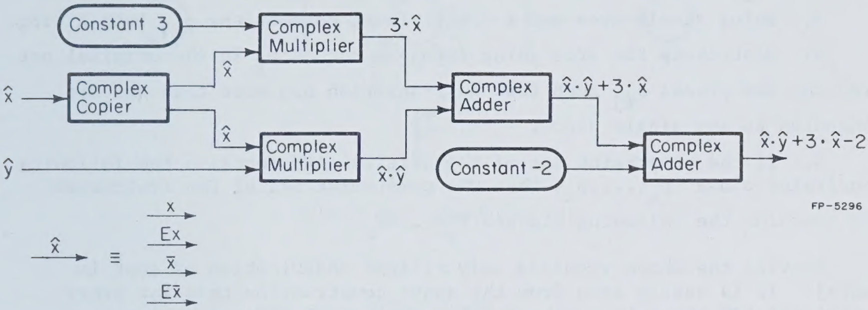


Figure 6. Example Polynomial Computer.

Figure 6. Calling this place p+ implies $b(p+,t_1) = 3$. Similarly for constant -2, except that one of the negative input places of the complex adder is used to introduce the constant.

Lemma 1: From the definition of a CSS together with the assertion at the beginning of this section and the construction method that follows it, it follows that;

$$C(CSS) \supseteq RES$$

where C(CSS) is the class of CSSs produced by all APNs.

Lemma 2: From the definition of an APN it should be clear that the class of all APNs can be enumerated. Hence the class of CSSs produced by all APNs can be enumerated, i.e.

$$RES \supseteq C(CSS)$$

Theorem: C(CSS) = RES follows from lemmas 1 and 2.

4. ON THE RELATIONSHIP BETWEEN APNs AND CNs

In the definition of APNs, we have allowed the use of multiple arcs to connect a particular transition and a particular place, i.e. the ranges of f and b are N. Nevertheless, for any APN with multiple arcs, it is possible to construct an APN with only a single arc connecting any transition and any place and which has the same CSS as the original APN. We will call such APNs *restricted* APNs. We say the two APNs are *behaviorally equivalent* since they have identical CSSs.

We now give a method for constructing a restricted APN which is behaviorally equivalent to a given APN. The method presented here is analogous to the one given by [Hack] for constructing a restricted Petri net which is equivalent to a generalized Petri net.

1. For each place, p_i, let k be the maximum number of arcs that go to or come from p_i.

2. Replace p_i with k new places, p_{i1}, p_{i2},... p_{ik}. Let this set of places be P_i.

3. Using single arcs and λ transitions, connect the p_{ij} into a ring.

4. Distribute the arcs going into and out of p_i in the original net over the new places p_{ij} such that no transition has more than one arc connected to any single place.

5. If the constraint set of the original APN contains the following constraint class $\{p_a,\ldots,p_z\}$ then the constraint set of the restricted APN contains the following classes $P_a \times \ldots \times P_z$.

Proving the above requires only a slight modification of that in [Hack]. It is easily seen from the above construction that for every restricted APN there is a behaviorally equivalent APN and vice versa.

The only difference between restricted APNs and CNs defined by Patil [Pati] is that a restricted APN has a special start place and special finish place. Peterson [Pete] effectively shows that the CSSs generated by Petri nets with start places and finish places are a subset of type 1 languages. This and the fact that the CSSs of restricted APNs must be the set of type 0 languages strongly suggest that the modelling power of CNs is stronger than that of Petri nets.

5. COROLLARIES

There are several corollaries of the Theorem in section 3 concerning the decidability of certain properties of an APN. The first such problem is reachability; that is, given an APN and a state, will the APN ever reach the given state? A second problem is boundedness; that is, for any particular place p, is there a finite number n such that the APN can never reach a state such that p has greater than n tokens? A final problem is coverability; that is, for a given APN and a given state, will the APN ever reach a state which has at least as many tokens in each place as the given state has?

Corollary 1: The reachability problem is undecidable.

Proof: The final state of an APN is only reached if the CSS of the APN is not emply. To determine emptiness of the CSS is equivalent to determining emptiness of a type 0 language. This problem is known to be undecidable.

Corollary 2: The boundedness problem is undecidable.

Proof: To the final place, F, append the places and transitions shown in Figure 7. The place F' becomes the new final place. We see that the place B becomes unbounded if F receives a token. Since it is undecidable whether F receives a token (see Corollary 1) it must be undecidable whether B is unbounded.

Corallary 3: The coverability problem is undecidable.

Proof: It is undecidable whether the final state is coverable.

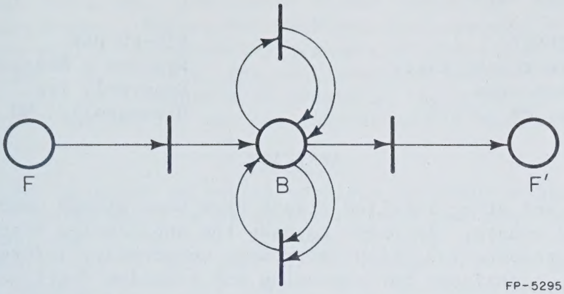


Figure 7. Corollary 2.

6. CONCLUSION

Although the behavior of APNs characterized by their CSSs, allows a greater modelling power (they can model any coordination of events which can be algorithmically described), the undecidability of the corollaries of section 5 rules out any universal techniques for answering some important questions concerning the computation processes that they might model. Since the validity of these corollaries has yet to be established for Petri nets, the same negative conclusions may not apply.

Since submitting this paper the authors became aware of similar prior work by Agerwala [Ager]. This proved the result of section 3 using a similar Petri net type machine called an Extended Petri Net. The result was achieved by showing that any Turing machine could be directly constructed as an Extended Petri net. This was also the original proof, used in this paper. However, so that this paper was not wholly replicative section 3 was redrafted with a different proof based on the results in [Mati].

7. REFERENCES

Ager Agerwala, T., "Towards a Theory for the Analysis and Synthesis of Systems Exhibiting Concurrency," Ph.D. Thesis, The Johns Hopkins Univ., 1975.

Hack Hack, M., "Decision Problems for Petri Nets and Vector Addition Systems," Technical Memo 59, Project MAC, MIT, Mar. 1975.

Mati Matijesavic, J. V., "Enumerable Sets are Diophantine," Soviet Math. Dokl., Vol. 11, No. 2 (1970), pp. 354-357.

Pati Patil, S. S., "Coordination of Asynchronous Events," Ph.D. Thesis, Dept. E.E., MIT, Jun. 1970, [MAC-TR-72].

Pete Peterson, J. L., "Modelling of Parallel Systems," Ph.D. Thesis, Dept. E.E., Stanford Univ., Dec. 1973.

Petr Petri, C. A., "Kommunikation Mit Automaten," Ph.D. Thesis, Univ. of Bonn, Germany, (1962).

REDUCTION OF PETRI NETS*

LARRY L. KINNEY
Dept. of Electrical Engr.
Univ. of Minnesota
Minneapolis, MN 55455

YIH-WU HAN
Systems & Research Center
Honeywell Inc.
Minneapolis, MN

ABSTRACT

Petri nets and other labelled graphs have been widely used to model real systems and events. In order to make the application practicable, a hierarchical representation which suppresses unnecessary information is essential. Hence, techniques for expanding and reducing Petri nets are needed. This paper discusses the reduction of live and bounded Petri nets which possess the properties of consistency and invariancy.

I. INTRODUCTION

Petri nets and other similar labelled graphs have been used as a representation tool for modeling digital system concepts such as parallel processing [1,2,3] performance evaluation [4,5] design language [6,7,8] and fault tolerance [9]. A Petri net with a slight modification is considered to be as powerful as the Turing machine with respect to representing the coordination of asynchronous processes [10].

A detailed Petri net representation of an actual system usually contains an enormous number of places, tokens and edges. This makes the application of Petri nets very time consuming and economically infeasible. One approach to overcoming this problem is to hierarchically represent the system at various levels of detail and suppress unnecessary details at higher levels of abstraction. This concept is based on a set of net transformations where a smaller net can be expanded into a larger net at a lower level of detail, and vice-versa. In addition to a set of transformation rules, a systematic verification method is needed to determine whether the parent and transformed net exhibit the same properties.

In this paper, we discuss the reduction of a Petri net to a smaller Petri net, such that the smaller net preserves some properties of the original net. Section II discusses the reduction of nets. In order to reduce them systematically, we need to identify some desirable properties of the nets and to determine mechanically whether a net possesses these properties. The desirable properties, primarily consistency and invariance, and their relations with liveness, boundedness and transition firing ratios are summarized. The threshold number of a net is introduced in Section III. If the threshold number of a net can be determined, reduction becomes straightforward.

II. REDUCTION OF PETRI NETS

Consider a Petri net containing a subnet which interfaces with the rest of the net only through input arcs to transition t_i, the input transition of the subnet, and output arcs from the transition t_o, the output transition of the subnet. In general, we want to determine what modifications can be made to the subnet and still obtain an "equivalent" Petri net. First, we need to define what is meant by equivalence.

---

*This research was supported by ONR under contract N00014-75-C-0011.

Suppose that our primary interest is in the events of the system modeled by the Petri net. The events of the system are represented by the firing of transitions. Hence, for each possible sequence of events in the system, there will a corresponding sequence of transition firings which can occur in the net. Conversely, for each possible sequence of transition firings there will be a sequence of events in the system. (It is assumed that the initial marking for the net corresponds to the initial conditions of the system.)

If we were to insist on having complete information about all events in the system, then it would be necessary to have complete information about all the transitions in the net. However, if only certain events are of interest, then it may be that part of the transitions in the net are superfluous. For example, if in the subnet only the events associated with transitions t_i and t_o are of interest, then it may be possible to reduce the subnet and eliminate some of the transitions within the subnet. To reduce the net, we should be able to handle the following issues:

(i) What properties must be preserved when making the reduction?

(ii) How can we determine whether a Petri net has these properties?

For purposes of discussing these reductions, we will use the following definition of a subnet.

Definition:

Let $N = \{P,T,A\}$ be a Petri net where P is the set of places, T is the set of transitions and A is a relation, $A \subseteq [P \times T] \cup [T \times P]$. A transition-to-transition subnet of N is the five-tuple $SN = \{SP,ST,SA,t_i,t_o\}$ where $SP \subseteq P$, $ST \subseteq T$, $SA \subseteq A$ and

(i) if t_j is in ST, $t_j \neq t_i$, then (t_j,p_k) in A implies that (t_j,p_k) is in SA,

(ii) if t_j is in ST, $t_j \neq t_o$, then (p_k,t_j) in A implies that (p_k,t_j) is in SA,

(iii) there is at least one p_k, such that (p_k,t_i) is in A but not in SA, and

(iv) there is at least one p_k such that (t_o,p_k) is in A but not in SA.

This definition simply says that a transition-to-transition subnet of Petri net consists of a subnet of the net's places and transitions and all of the connections between these places and transitions except for some input edges to t_i and some output edges from t_o. We refer to t_i as the input transition of the subnet and t_o as the output transition. In an analogous fashion we can define transition-to-place, place-to-transition, and place-to-place subnets. Unless otherwise indicated, subnet will mean transition-to-transition subnet.

The reduction problem can now be stated as follows. Let $N = \{P,T,A\}$ be Petri net and $SN = \{SP,ST,SA,t_i,t_o\}$ be a subnet of N. The subnet SN is reduced to a new subnet $SN^1 = \{SP^1,ST^1,SA^1,t_i,t_o\}$ where $SP^1 \subseteq SP$, $ST^1 \subseteq ST$, $SA^1 \subseteq SA$, t_i is in ST^1 and t_o is in ST^1. The rest of N combined with SN^1 produces a reduced net N^1. Then, N with initial marking M is equivalent with respect to transitions t_i and t_o to N^1 if there is an initial marking M^1 for N^1 such that the set of all possible transition firings of t_i and t_o are the same for N and N^1.

For some simple nets, it is possible to reduce a subnet by inspection. For example, in the net of Figure 1a, let t_i, t_1 and t_o be transitions in the subnet. Clearly, the only possible sequence of firings of t_i and t_o is $t_i, t_o, t_i, t_o, \ldots.$ It is also obvious that this is the only possible sequence of firings of t_i and t_o for the net of Figure 1b, hence, these two nets are equivalent with respect to t_i and t_o.

As another example of a reduction, consider Figures 1c and 1d. These nets are equivalent and the possible firing sequences are those beginning with t_i in which t_i has fired 0, 1, or 2 more times than t_o. These examples illustrate an important point concerning reduction of subnets. Note that the subnets of Figures 1a and 1c are identical; however, the reduced subnet of Figure 1b is not a valid reduction for the subnet of Figure 1c. The difference between the two cases is determined by the portion of the net external to the subnet -- in this case, a single place with tokens.

It restricts the number of times that the input transition t_i can fire before the output transition t_o must fire. For Figure 1a, the number is 1 while for Figure 1c, it is 3. Clearly, the reduced subnet of Figure 1d is valid no matter how many times t_i can fire. Furthermore, the reduction of this subnet can be made independent of the portion of net external to subnet.

In the remainder of this paper, subnets like the preceding one will not be considered. Only "well-behaved" subnets will be considered. We define a net to be "well-behaved" if it possesses the properties of consistency and invariance. These properties are defined below. This paper will also be confined to the reduction of transition-to-transition subnets. It is, of course, possible to consider the reduction of other types of subnets.

There are other examples of simple series and parallel connections of transitions where the reduction can be made by inspection. In some the output transition can fire before the input transition fires. This is analogous to the transient response of systems. In some cases, a net can exhibit a very "nonlinear" transient behavior. It is easy to construct an example where the output transition t_o can continue to fire indefinitely until the input transition t_i is fired. However, once the input transition fires, the output transition can fire one more time and the net is dead thereafter.

In the remainder of this paper, subnets like the preceding one will not be considered. Only "well-behaved" subnets will be considered. We define a net to be "well-behaved" if it possesses consistency and invariance. These properties are defined below. This paper will also be confined to the reduction of transition-to-transition subnets. It is, of course, possible to consider the reduction of other types of subnets.

Suppose that to every transition in a net, a variable is assigned. Let the variables be constrained by requiring that, for every place, the sum of the variables assigned to input places be equal to the sum of the variables assigned to output places. If there exists a positive integer (nonzero) solution for the variable, then the net is called consistent. If a net is consistent, then a consistent solution specifies the number of transition firings which occur for some cyclic firing sequence and conversely. (A cyclic firing sequence is a sequence of transition firings which fires each transition and has the same final marking as initial marking.)

The property of invariancy is defined in a dual manner. Variables are assigned to places and are constrained by requiring that, for every transition, the sum of the variables assigned to input places is equal to the sum of the variables assigned to output places. If there exists a positive, integer (nonzero) solution for the variables, the net is called invariant. Invariancy is a sufficient, but not necessary, condition for a net to possess a bounded marking. There are example of nets which are both consistent and invariant and other examples which are neither. Also, there are examples which are one and not the other.

III. THE THRESHOLD NUMBER OF A PETRI NET

In most Petri nets which represent real systems, the maximum number of firings of a specified input transition without firing a specified output transition is fixed. This number is defined as the threshold number with respect to the input transition and the output transition. It is obviously controlled by the number of tokens in the paths leading from the output transition to the input transition. These paths are called feedback paths in contrast to the paths leading from the input transition to the output transition which are called forward paths. Let us also define a simple path as a path which does not cover any transition or any place more than once. Then, rigorously, a feedback path is a simple path from t_o to t_i; and a forward path is a simple path from t_i to t_o.

In the discussion of threshold number, it is assumed there exist feedback paths to control the token population in nets concerned, since otherwise there exists no threshold number; i.e., it is undefined or infinite. The scope of discussion is on reducible nets only, which are defined as follows. If a Petri net between an input transition and an output transition can be reduced to a net as shown in Figure 2, formed by a single loop consisting of a forward path and a feedback path, and given an initial marking, the token number in the feedback path is a fixed positive integer, then this net is called completely reducible or reducible for short.

Note that if a net is reducible then the firing ratio between the input and output transition is one-to-one. This assumes that a cyclic firing sequence is being considered; i.e., a firing sequence which returns the net to the same initial conditions. In addition, the maximum number of times that the input transition can fire before the output transition must fire is m, where m is the threshold number. This specifies the "transient behavior" of the net. As discussed previously, only live, consistent, invariant Petri nets are considered. Hence, a completely reducible net

- Is live
- Is consistent
- Is invariant
- Has a 1 to 1 firing ratio between input and output transitions (for cyclic firing sequences)
- Has a positive, constant threshold number

If a net is given, it is a relatively easy task to check for consistency and invariancy and, thereby, check for a 1 to 1 firing ratio between input and output transitions. If a simple procedure is available to decide if the net has a positive constant, threshold number and, if so, to determine its value, then reduction of the net is straightforward. However, the determination of the threshold number appears to be difficult for general reducible nets. The approach investigated to date is to examine the type of coupling between forward and feedback paths and between feedback paths.

The threshold number is related to the minimum number of tokens which can exist on feedback paths. When the input transition fires, a token is removed from each feedback path and a token is added to each forward path. When the output transition fires, the reverse effect takes place. Clearly then, the maximum number of times which the input transition can fire without firing the output transition is determined by the minimum number of tokens which can exist on a feedback path. (This is true provided there is no self-loop between the input transition and a place on the feedback path.) However, this number is difficult to determine because the firing of internal transitions in the subnet can increase or decrease the number of tokens on a feedback path. The coupling between forward paths and feedback paths and between feedback paths are discussed below.

Consider the simple subnet of Figure 3. The feedback path and the forward path are coupled at a place. The firing of the internal transition can remove a token from the feedback path which is later returned by the firing of the output transition. The net has an indeterminate threshold number. The input transition can fire an arbitrarily large number of times without firing the output transition and, conversely, the output transition can fire an arbitrarily large number of times without firing the input transition.

The preceding example shows that the coupling of feedback and forward paths at places can cause the threshold number to be undefined. However, this is not necessarily true. For example, in Figure 4 the same type of coupling exists but, nevertheless, the threshold number is well defined (it is 1). In contrast to Figure 3, the net of Figure 4 is consistent with a unique solution for the assignment of values to the input and output transitions. Although place p in Figure 4 has two output transitions t_i and t_1, they are controlled by a loop with only one token in t_i. Thus, only one of t_i or t_1 can fire when p contains a token. With this kind of coupling, the procedure for determining the threshold number of a Petri net is very involved. For simplicity, we shall only deal with Petri nets in which the type of coupling between any forward path and any feedback path occurs only at transitions.

If there is a feedback path from t_o to t_i which does not couple with any other paths at places, then the only way to place more tokens on that path is to fire the output transition t_o. Hence, clearly if m is the number of tokens on the feedback path, the input transition t_i cannot fire more than m times without firing t_o. If there are two or more feedback paths with coupling at places, then the situation is more complicated. For example, if two feedback paths are coupled from a place p_2 to a place p_1, then all of the tokens can be moved to place p_1. Thus, t_i cannot fire more times than the sum of the tokens on the two paths without firing t_o.

If two feedback paths are coupled from a place to a transition, then the net may not be persistent and the number of times which t_i can fire is not defined; i.e., it is nondeterministic. However, this is not necessarily true for every case of this type of coupling.

A third type of coupling between feedback paths is transition-to-place. This case is similar to place-to-place coupling. Obviously, the place where the two feedback paths join can receive at most a number of tokens equal to the sum of the number of tokens on the portions of the two paths from t_o to the place. (Note that the tokens on the initial overlapping portion of the paths are counted twice -- once for each path.) If we add to this, the number of tokens on the two paths between the place and t_i this gives the number of times t_i can fire without firing t_o. (Note that the tokens on the final overlapping portion of the paths are only counted once.) Again this limit is achievable provided some other feedback path does not have a lower limit.

Finally, the last type of coupling which can occur between feedback paths is transition-to- transition. Clearly, in this case, the transition where the paths join cannot fire more times than the minimum of the number of tokens on the two paths from t_o to that transition. Then, it also follows that t_i cannot fire more than the minimum number of tokens on the feedback paths.

All types of coupling between feedback paths, namely place-to-place, place-to-transition, transition-to-place and transition-to-transition, have been intuitively discussed. Again, for simplicity, we shall deal with only the transition-to-transition coupling between feedback paths. For these restricted Petri nets, the following lemma and theorem illustrate the way to determine threshold numbers. In the discussion, internal transitions of a subnet are all the transitions of the subnet except t_i and t_o.

Lemma 2

Let $\{P,T,A,t_i,t_o\}$ be a reducible subnet with only transition-to-transition coupling between feedback paths and between forward and feedback paths. If transition t_i is disabled and can only be enabled by the firing of t_o, then there exists a feedback path which is free of tokens in the subnet.

Proof:

Since t_i is disabled, it has at least one input place which is free of tokens. We want to show that there is at least one such place, which remains free of tokens no matter what internal transitions fire. If this were not true, there would be two places, say p_1 and p_2, such that a token can be on one or the other but not both. (The argument easily extends if there is more than two places.) However, this implies that there is a path from p_1 to p_2 and also a path from p_2 to p_1. Either p_1 or p_2 or both must be on a feedback path; otherwise, the firing of t_o could never cause the placement of a token on them and the subsequent enabling of t_i. If both p_1 and p_2 are on feedback paths, then these feedback paths are coupled at places. It is a contradiction. Similarly, if one is on a forward path and the other on a feedback path then the forward path and the feedback path are coupled at places. Again, it is a contradiction. We conclude that there is one input place, say p_1, to t_i which remains free of tokens until after t_o fires.

Let t_1 be the input transition to $p_1 \cdot t_1$ remains disabled until after t_o fires. Thus, t_1 always has one input place free of tokens. By the same argument as above, there is at least one place, say p_2, which remains free of tokens until after t_o fires.

Continuing in this way, a feedback path $t_i p_1 t_1 p_2 t_2$ ... can be constructed which is free of tokens. Q.E.D.

Let us define the token content of a feedback path as the sum of tokens in the places of the path. The minimum token content of the feedback paths or just minimum token content for short is the minimum value of the token contents. The following theorem says that for a group of reducible Petri nets specified previously, the threshold number is equal to the minimum token content.

Theorem 4:

Given a reducible net with only transition-to-transition coupling between feedback paths and between forward and feedback paths, the threshold number, m, is equal to the minimum token content, f, of the feedback paths.

Proof:

(i) $m \leq f$

This can be proved easily from the first two of the following observations:

1. Every firing of t_i removes a token from every feedback path.
2. Every firing of an internal transition does not change the token content of any feedback path. This is true, because only transition-to-transition coupling is considered.
3. Every firing of t_o adds a token to every feedback path. (This condition is not needed for this part of proof, it is mentioned for completeness.) Thus, by firing t_i f times without firing t_o,we shall find that at least one feedback path is free of tokens. Therefore, t_i cannot fire again until after t_o fires. So, $m \leq f$.

(ii) $m \geq f$

The proof is by contradiction, suppose $m<f$. That means after firing t_i m times without firing t_o, t_i cannot fire any more until the firing of t_o. From observations 1 and 2, and the definition of the minimum token content, it is clear that every feedback path should contain at least one token. But from Lemma 2, there exists a path free of tokens. This is a contradiction. Therefore, $m \geq f$.

It has been shown that $m \leq f$ and $m \geq f$, so $m=f$. Q.E.D. Obtaining the threshold number, m, of a net enables us to reduce the net in the form shown in Figure 2.

V. CONCLUSIONS

The study of this paper is motivated by the goal to establish a mathematical foundation for a hierarchical representation, which suppresses unnecessary information and allows later expansion, if necessary. Although only the reduction of a net has been discussed, the techniques for determining the consistency, invariance, transition firing ratios and threshold number of a net also provide a means to some extent for verifying the correctness of expansion. Because after expanding a net, we also can examine whether these properties are preserved. The scope of the paper will be summarized below.

Consistency and invariance have been defined. Their physical meanings are suggested. A scheme for verifying that a net possesses the properties of consistency and invariance has been proposed. By solving the homogeneous consistency equations, the transition firing ratios are also obtained. For a reducible net with only the transition-to-transition coupling, we are able to determine the threshold number, and thus reduce the net in the form shown in Figure 2.

There are still several problems remaining to be solved. For example, the scheme for determining the threshold number should be generalized to less restricted Petri nets. A mechanical way is required to select the input transitions for a subnet. In addition, the procedures for reducing a subnet with input places and output places has not been tackled yet.

REFERENCES

1. R. M. Karp and R. M. Miller, "Properties of a Model for Parallel Computations: Determinacy, Termination, Queuing", SIAM J. Appl. Math., November 1966, pp. 1390-1411.

2. R. M. Karp and R. E. Miller, "Parallel Program Schemata", J. of Computer & System Sciences, 1969, pp. 147-195.

3. J. D. Noe and G. J. Nutt, "Macro E-Nets for Representation of Parallel Systems", IEEE Trans. on Computers, Vol. C-22, No. 8, August 1973, pp. 718-727.

4. G. J. Nutt, and J. D. Noe, "Validation of a Trace Driven CDC 6400 Simulation", Proc. 1972 SJGG, Vol. 40, pp. 749-757.

5. G. J. Nutt, "Evaluation Nets for Computer System Performance Evaluation", University of Washington, TR #72-04-03, April 1972.

6. J. B. Dennis, "First Version of a Data Flow Procedure Language", Project MAC, MIT Group Memo 93-1, August 1974.

7. C. W. Rose and M. Albarran, "Modeling and Design Description of Hierarchical Hardware/Software Systems", Proceedings of 12th Design Automation Conference, 1975.

8. M. Hack, "Petri Net Languages", Lab. for Computer Science, MIT, RT-159, March 1976.

9. L. A. Jack, W. L. Heimerdinger and M. D. Johnson, "Theory of Fault Tolerance", Honeywell Annual Report, System & Research Center, Honeywell, Inc., September 1975.

10. T. Agerwala, "Towards a Theory for the Analysis and Synthesis of Systems Exhibiting Concurrency", Ph.D. Thesis, Johns Hopkins University, 1975.

11. F. Furtek, "A New Approach to Petri Nets", Project MAC, MIT, Group Memo 123, April 1975.

12. C. Ramchandani, "Analysis of Asynchronous Concurrent Systems by Petri Nets", Project MAC, MIT, TR-120, Feburary 1974.

13. A. W. Holt and F. Commoner, "Events and Conditions: An Approach to the Description and Analysis of Dynamic Systems", Third Semi-Annual Technical Report, Part II, for the Project "Research in Machine-Independent Software Programming", Applied Data Research, Inc., April 1970.

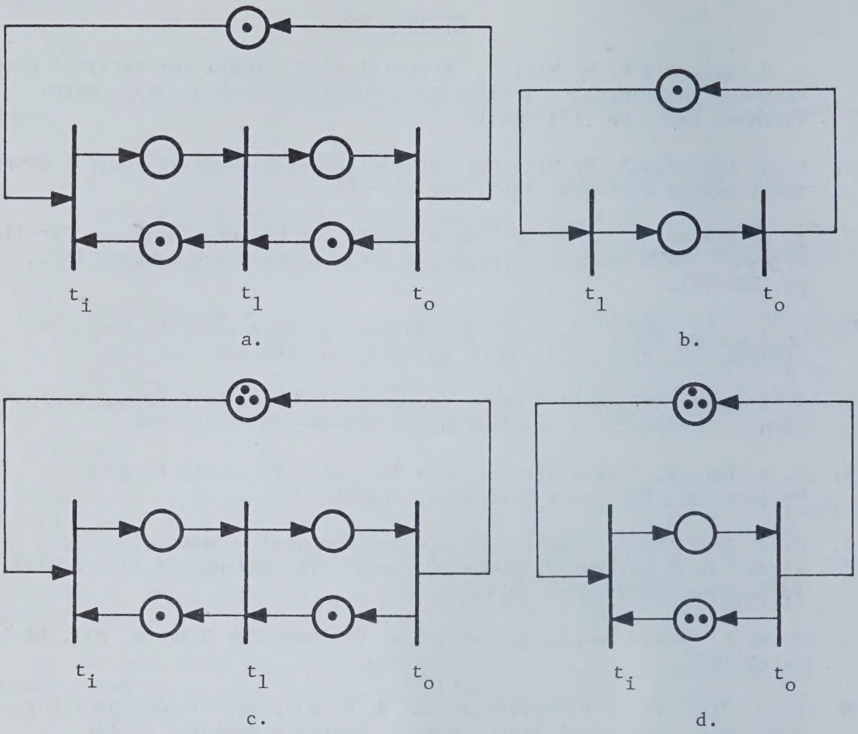
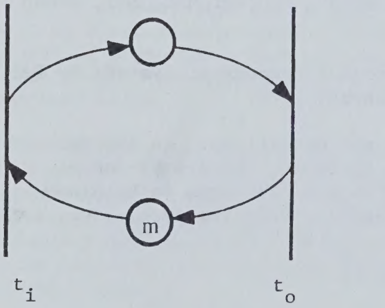

t_i
t_1
t_o
a.
t_1
t_o
b.
t_i
t_1
t_o
c.
t_i
t_o
d.

Figure 1. Reductions Which Depend Upon the External Net



where m is the threshold number

Figure 2. Reducible Net

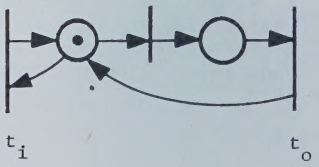


Figure 3. A Net Which has Undefined Threshold Number

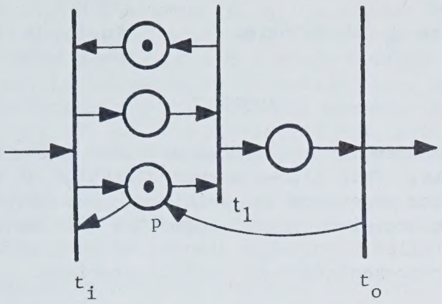


Figure 4. A Net With a Defined Threshold Number

THE CONTROL PROBLEM FOR SEQUENTIAL MACHINES

DAVID W. FRALEY
Battelle Northwest Laboratories
Richland, WA 99352

JOSEPH J. TALAVAGE
Purdue University
West Lafayette, IN 47905

ABSTRACT

The control problem for sequential machines is defined and extended from present results. This allows a generalization of the process of defining a control problem on sequential machines. The control problem is given as a discrete decision process augmented with the structure of the machine to be controlled. Solvable classes of such problems are addressed in terms of their representation as control problems.

I. INTRODUCTION

In systems engineering the emphasis is upon problem solving within a given framework. The elements of systems engineering problems, particularly those dealing with systems control, have been well stated by Sage and Melsa [1], as: a goal or objective, determination of position with respect to the goal (a measure of effectiveness of performance); determination of the environmental factors affecting the system and the establishment of a system model (identification); and determination of a policy in accordance with goal definition, current state and the model itself (computation).

These intuitive notions have been applied extensively to continuous-state systems and the body of knowledge, or discipline, is often referred to as control theory. Whereas control theory is pragmatic in its techniques, the extant knowledge in automata theory emphasizes its algebraic aspects. In addition, the more esoteric questions about the process of problem solving itself, or of computation, are often considered to be within the realm of automata theory. To elucidate the similarities between control theory and automata, we emphasize the practical aspects from both. The heart of such an outlook is the system itself.

Consider a system M, which is time-invariant as a quintuple -- $M = (Q, \Omega, Y, \delta, \lambda)$, where Q is a set of states, Ω is a set of acceptable input segments, Y is a set of outputs, $\delta : Q \chi \Omega \rightarrow Q$ = state transition function, $\lambda : Q \chi \Omega \rightarrow Y$ = output function, (where χ denotes Cartesian product).

Although time is not explicit, this quintuple can be interpreted as a system which if at time t_o is in state q and receives input segment ω from time t_o to t, will at time t be in state $\delta(q,\omega)$ and will emit output $\lambda (q,\omega)$. This system becomes an automaton in the formal sense if the input and output sets are finite and time is discrete. On a discrete time scale the set of admissible input segments becomes a set of finite sequences on the finite input set X. Thus the set X determines Ω. The set of all finite sequences of the elements of the set X; X*, which includes the empty sequence Λ may be considered. Furthermore, if Q has finitely many members, then M is a finite automaton or finite state machine.

Definition 1. A sequential machine is a system $M = (Q, X, Y, \delta, \lambda)$ where Q, X, Y, δ, and λ are given in the above discussion.

We shall say that a sequential machine M is finite when Q, X, and Y are finite sets. In some cases [2], alternative definitions include the designation of a state $q_o \in Q$ denoted (M, q_o) and a set of states $F \subseteq Q$, known respectively as the initial state and the set of final states. The mappings δ and λ are from $Q \chi X$ to Q and Y respectively. To consider input segments, or sentences, it is necessary to extend these mappings to the domain $Q \chi X^*$. This is done in the conventional manner: $\delta : Q \chi X^* \rightarrow Q$ such that for each $q \in Q$, $x \in X^*$, and $a \in X$, $\delta(q, \Lambda) = q$, and $\delta(q, xa) = \delta(\delta(q, x), a)$; $\lambda : Q \chi X^* \rightarrow Y$ such that for each $q \in Q$, $x \in X^*$, and $a \in X$, $\lambda(q, xa) = \lambda(\delta(q,x),a)$. A sentence $x \in X^*$ is said to be accepted by M if $\delta(q_o, x) = q$ for some $q \in F$, where M has defined an initial state and set of final states. In this case, the system of our concern is a sequential machine.

A discrete optimization problem, given in the form of a discrete decision process is augmented with a sequential machine to form a control problem. The resultant control problem is referred to as a sequential decision process; a formal structure for the application of dynamic programming. Dynamic programming is a technique or algorithm (in the sense of computation) to obtain the set of optimal solutions for many discrete optimization problems.

Dynamic programming as characterized by Bellman [3] through the "principle of optimality", has become an important technique for solving problems involving sequential processes. Many view the sequential process as one which evolves from state to state in response to decisions - a sequential decision process. Since the appearance of Bellman's work many investigators have been providing a rigorous mathematical basis for the "principle of optimality". Notable are the works of Mitten [4], Nemhauser [5], Denardo [6] and others. Another indication of the principle's utility in dealing with sequential systems has been the power it has exhibited in solving problems from linear systems theory.

In particular, Karp and Held [7] showed the relationship of a discrete decision process as a means of problem statement to a sequential decision process as a structure to which dynamic programming may be applied. The authors above made clear the role of a basic monotonicity property of cost functions in justification of the functional equations of dynamic programming. Cost functions provide a measure of performance.

How does all of this relate to sequential machines? It turns out, as shown by Karp and Held, that monotone sequential decision processes may be characterized as finite sequential machines with a certain cost structure superimposed. Ibaraki [8-10] has refined that concept and defined various subclasses of monotone sequential decision processes. With this set of models, Ibaraki has further addressed decision problems in the sense of the theory of computability; i.e., whether dynamic programming is applicable to a given sequential decision process. In this light he has treated such processes as recursive functions and given various subclasses for which there exist algorithms or solvable problems.

Here a unique point-of-view is taken. Because a sequential decision process is representable by a sequential machine augmented with a discrete decision process, then the control problem for a sequential machine may be represented by a sequential decision process. The decision process by which we ultimately control it, if possible, will be based on the structure of the machine itself. This, in turn, is determined by the model that we utilize.

In this paper, the control problem for sequential machines is defined. It is a sequential machine with a cost function associated with each state transition - a sequential decision process. To ensure the existence of solutions to control problems for sequential machines, the process is restricted to be recursive and in most cases to satisfy certain monotonicity conditions. With these restrictions three subclasses of sequential decision processes are introduced. For all three subclasses, the functional equations of dynamic programming apply and procedures to obtain the optimal solutions exist.

Section II generalizes the definition of cost space and subsequently defines the control problem for sequential machines. Background is then provided on discrete and sequential decision processes in the third section. The properties of recursiveness and monotonicity are given to restrict the classes of sequential decision processes applicable as control problems. Such restrictions are related to the existence of solutions to the control problem, as developed by others in the form of solutions to classes of sequential decision processes.

The fourth section characterizes the solvable classes of sequential decision processes as to the three abovementioned "types". Then the augmentation of discrete decision processes with machines to form a control problem is given in Section V. The final section develops some of the structural properties implicit in the machine and its related cost function which should be considered in defining a control problem on sequential machines.

II. THE CONTROL PROBLEM

Generally, modern control theory has been concerned with either discrete or continuous systems which can be described by differential or difference equations, generally linear. As a result various solution techniques have been able to capitalize on the continuity of such systems since they are defined on the vector spaces of the real numbers. Then it makes sense to speak of performance to within some desired margin. To alleviate this problem for sequential machines, Arbib [11] has defined a tolerance.

Definition 2. A tolerance ζ on a set Σ is a relation that is reflexive and symmetric. A tolerance space is the set together with its tolerance(Σ,ζ).

With the concept of tolerance it makes sense to speak of performance for sequential machines. For example, a tolerance space may be the set of states reachable in a given interval of steps. Let $\{\sigma j(q)\}$ be the set of all states reachable from state q in at most j steps. Those states reachable in $k<j$ steps, such that $\{\sigma j(q)\}$ is reachable in $|j-k|$ steps, is a tolerance on the set of desired states.

In order to evaluate performance, relative to a goal, we must assign a performance measure. Generally this is referred to as a cost function. Arbib's [12] definition is extended as follows:

Definition 3. A cost space C is a tolerance space for some ζ which is ordered by $\leq$ and for every c, $c' \varepsilon$ C there exists a, b ε C such that $a<c<b$ and $a<c'<b;(c,c') \varepsilon \zeta$. Given a machine $M = (Q, X, Y, \delta, \lambda)$ and a cost space C, a cost function for M is a map $h : C \chi Q \chi X \rightarrow C$.

With a function which assigns a measure of effectiveness or performance we may then define the objective or goal as the control problem.

Definition 4. Control problem. Given a sequential machine (M,q_o), and cost function h, let $q \varepsilon F$ be a state in the set of final states of M. The sequence of input symbols $x \varepsilon X^*$ which transfers M from initial state q_o to state $q \varepsilon F$ if $\delta(q_o,x) = q$ is a control. If we determine all such sequences which achieve this transfer to any $q \varepsilon F$; find $x \varepsilon X^*$ which minimizes the associated cost function.

Before considering the control problem for sequential machines in any detail, we consider sequential decision processes, so we may base our discussions in the proper perspective.

III. DISCRETE AND SEQUENTIAL DECISION PROCESSES

The control problem for sequential machines can be viewed as an optimization problem involving a sequence of decisions. In other words, the cost space along with the possible decisions $(x \varepsilon X)$ and strings of decisions or policies $(S \subset X^*)$ can be given in the form of a discrete decision process. If such a space is associated with a sequential machine, then the decisions are elements of the input alphabet for that machine and the strings $x \varepsilon X^*$ in M are policies. A machine M plus a cost function form a sequential decision process. Then the solution of the control problem for sequential machines is equivalent to determining the set of optimal policies for a sequential decision process.

Dynamic programming is an important technique for the solution of problems involving the optimization of a sequence of decisions. The simple idea underlying this technique is to represent the problem by a process which evolves from state to state in response to decisions. A typical optimization problem then becomes one of guiding the system to a terminal configuration at minimum cost. Then, when the cost structure is appropriate, the determination of the optimal policy, e.g. sequence of decisions, may be reduced to the solution of a functional equation in which the state appears as an independent variable. The main elements of such a process are a set of primitive decisions, a set of policies, and a cost function on the policies.

Definition 5. A discrete decision process $T(X,S,h)$ is defined as: X is an alphabet (primitive decisions). X^* is the set of strings (policies) obtained by concatenating finite primitive decisions. $S \subset X^*$ is a set of feasible policies. $h : S \to C$ is a cost function.

We generally think of a cost space as the real numbers under a tolerance of the form $(c',c) \varepsilon \zeta \Rightarrow |c-c'| < \nu$ for some $\nu>0$. A policy x is feasible if $x \varepsilon S$ and optimal if $x \varepsilon S$ and for every $y \varepsilon S$; $h(x) \leq h(y)$.

A sequential decision process is more structured. Such a process evolves by transitions and cost is accumulated by increments in response to decisions. Discrete decision processes may be related to input-output behavior of sequential machines; sequential decision processes relate to the internal structure of the machine.

Associated with each sequential decision process is a unique discrete decision process. Furthermore, if the sequential decision process has a certain monotonicity property, as given by an appropriate cost space, then a set of recurrence relations can be given which determines the minimum cost policies for the discrete decision process. These recurrence relations are the functional equations of dynamic programming. The dynamic programming approach to solving the minimization problem for a discrete decision process depends upon a representation of the process as a sequential decision process.

Definition 6. A sequential decision process Π is a system $\Pi = (M, h, c_o)$ where: M is a sequential machine $M(Q,X,Y,\delta,\lambda)$ with an initial state q_o and a set of final states $F \subset Q$; $h: C \times Q \times X \to C$ is a cost function; and $c_o \in R$ where R is a real number and c_o is an initial cost value for q_o.

The real-valued function h is used to introduce the notion of "cost" into the sequential decision process. The quantity $h(c,q,x)$ gives the cost of reaching the state $\delta(q,x)$ by an input that reaches q at cost c. The map h can be extended as with finite machines to $C \times Q \times X^* \to C$ by $h(c,q,xx') = h(h(c,q,x), \delta(q,x),x')$ with $x,x' \in X$. Similarly, the quantity $h(c_o,q_o,x)$ gives the cost of starting in an initial state q_o at cost c_o and then applying the input sequence $x \in X^*$.

In a sequential decision process the cost is accumulated step by step whereas in a discrete decision process the cost may be presented in an arbitrary manner. The minization problem for such a process is to construct an algorithm which given any h minimizes $h(c_o,q_o,x)$ over the set of sentences accepted by the finite sequential machine M. Note that by associating the structure of a machine in our definition we can truly address the control problem from the point of view of a decision process.

In effect, the control problem for sequential machines is a minimization problem for a sequential decision process. We may restate Definition 4 as: Construct an algorithm which, given an h, minimizes $h(x)$ over all $x \in X^*$ for a sequential decision process $\Pi = (M,h,c_o)$.

A sequential decision process Π is said to represent the discrete decision process T if $S \subseteq X^*$ is the set of words accepted by M and their cost functions are identical.

If the sequential machine is a finite one, then we can address computability again in terms of recursive processes. A recursive discrete decision process T is a discrete decision process with additional restrictions that: S is regular, i.e. S can be accepted by some finite sequential machine; $h(x)$ takes on only integral values for $x \in S$; and $h : X^* \to C$ where $C \subset Z$ with Z the set of integers.

Here then, h is a partial recursive function or in other words is effectively computable. A partial recursive function is a function whose value can be evolved in a finite number of steps for any argument in its domain. This implies that the domain of h is the equivalence relation defined on X^* such that the set of x with $h(x)$ defined is the set S. That is, the equivalence class $\{x | h(x) \text{ defined}\}$ equals S.

Similarly a recursive sequential decision process (r-sdp) Π has the restrictions: $C \subset Z$; and the domain of $h \subset L_\pi$ where

$$L_\pi = \{(h(c_o,q_o,x), \delta(q_o,x),x') \mid x \in X^*, x' \in X\}.$$

Here, as above, the cost functional takes on only discrete values, i.e., $h : Z \times Q \times X \to Z$ and its values are in the class defined by L_π. That is, L_π is such that $h(c_o,q_o,x)$ and $\delta(q_o,x)$ are defined and the policy x' is in X_π, then h takes its values from the costs provided by the structure of the cost function and the transitions of the machine M. By the restriction that such processes by recursive, we may address those optimization problems which can be effectively computed.

Next, we simplify the minimization problem so that it can be characterized as the solution to a set of recurrence relations of the type normally associated with dynamic programming. To do this we restrict the class of sequential decision processes under consideration to those which possess a certain monotonicity property. That is, the sequential decision process Π is said to be monotone if $c \leq c'$ implies $h(c,q,x) \leq h(c',q,x)$ for every $c,c' \varepsilon C$, $q \varepsilon Q$ and $x \varepsilon X^*$. Furthermore, the process is both monotone and recursive if (c,q,x) and (c',q,x) are restricted to L_π.

If a sequential decision process is both monotone and recursive, it follows that the functional equations of dynamic programming represent a technique for solving the control problem for sequential machines. First, a set of feasible policies are given as the set accepted by the finite automaton. A set of optimal policies are such that if a string of inputs $x \varepsilon X^*$ are feasible then the set

$$O(\Pi) = \{\{ x \mid \delta(q_o,x) \varepsilon F\} \mid (\text{For every } y \mid \delta(q_o,y) \varepsilon F) \\ (h(c_o,q_o,x) \leq h(c_o,q_o,y))\}$$

is a set of optimal policies. From Definition 4, we see that a control is a feasible policy and the above set $O(\Pi)$ yields a restatement of the definition.

It was first shown by Karp and Held that the functional equations of dynamic programming always hold for a monotone sequential decision process. By restricting the process to be recursive (e.g. computable) Ibaraki [8] restated that result. That is;

$$H(q_o) = \min\ [c_o, \min\{h(H(q'),q',x) \mid \delta(q',x) = q_o\}],$$

$$H(q\ ) = \min\ \{h(H(q'),q',x) \mid \delta(q',x) = q\}\ q \neq q_o$$

where $H(q) = \inf\ \{h(c_o,q_o,x) \mid \delta(q_o,x) = q\}\ q \varepsilon Q$, (inf=g.l.b)

and for every $q' \varepsilon Q$, $x \varepsilon X^*$; $h(-\infty,q',x) = -\infty$,

where $-\infty$ is a special symbol smaller than any tolerance defining the cost space. If the functional equations $H(q_o)$ are solvable then $\min\ \{H(q) \mid q \varepsilon F\}$ gives the value of the optimal policies. Thus a recursive-msdp may be considered as a general model for SDP's to which the method of dynamic programming is applicable.

Using this result it is possible to devise an approach to the control problem for sequential machines. The first step, given a machine, is to represent the control problem we wish to solve by a sequential decision process. The second step is to determine the representational equations above. Next is to give a method of solving these equations such that an optimal input sentence is determined. From this, our primary concern is to characterize the representations of sequential machines as sequential decision processes - particularly those which are monotone and recursive.

IV. SOME PROPERTIES OF SEQUENTIAL DECISION PROCESSES

In this section, three specific subclasses of sequential decision processes are considered. In each, optimal policies exist and can be determined. A fourth process is also mentioned; for this case the existence of optimal policies must be known apriori if it is to be utilized as a control problem.

In defining tolerances and costs for the characterization of the control problem for sequential machines, Arbib [12] restricted the cost space to be additive. Here we have extended this definition; the form of an additive process is considered. Our definition subsumes that previously given, since we consider an additive process a subset of one of the three classes to be considered.

It is easy to show that additive processes are monotone. The additive form can be given as; $h(c,q,x) = c+\ell(q,x)$, where $\ell(q,x)$ can be interpreted as the incremental cost of applying input x while in state q. The implication here is that such cost functions do not necessarily have to be additive to define a control problem for sequential machines. Intuitively, for any optimal policy if the process is additive, the costs incurred along the way are minimal among the alternatives. This result is similar to the principle of optimality.

The present results in this area deal with the characterization of discrete decision processes as sequential decision processes. For example, does there exist an algorithm to decide whether or not an arbitrary discrete decision process can be represented by a sequential decision process? Here, we are concerned with a reverse question - what sort of discrete decision process is applicable to the given sequential machine?

We are assuming that we have a sequential machine and some arbitrary discrete decision process which we have defined on that machine. Hence, we have a sequential decision process. Then our concern really is whether or not an optimal policy can be determined.

In practical terms, we desire to characterize a control problem for a sequential machine by a sequential decision process. In the previous section, from the results of Ibaraki [13] and Karp and Held, such processes were restricted to be monotone. Then, since these models are still too general to discuss decision problems in the sense of computation [14] they were further restricted to be recursive. As it turns out, the question "does there exist an algorithm to obtain an optimal solution?" is still undecidable [9] at this generality. To get around this difficulty Ibaraki [9,14] introduced several classes of recursive-monotone sequential decision processes, each having an algorithm to obtain an optimal policy, provided that one exists.

Recall that a function is recursive, rather than partially recursive if the domain of the function is completely defined. A recursive sequential decision process is not partially recursive if $h(c,q,x)$ is defined for every $c \in Z$, $q \in Q$ and $x \in X$ even if there exists no $x \in X^*$ such that $\delta(q_o,x) = q$ and $h(c_o,q,x) = c$. In other words, costs are assigned for all transitions in our machines even though no feasible policy exists.

The first subclass considered by Ibaraki [14] is the recursive-invertible monotone sequential decision process (r-imsdp). Such a process has a distinguishing feature that there exists an algorithm to obtain an optimal policy $x \in O(\Pi)$ provided that $O(\Pi)$ is not empty or $O(\Pi) \neq \phi$. In practical applications, this property is quite important since there exists no such algorithm for general recursive-monotone sequential decision processes.

In our situation if we are given a control problem for a sequential machine, as a recursive-invertible monotone sequential decision process, and told that there exists an optimal policy, then a policy could be determined. On the other hand, if we wish to assign a decision process to a sequential machine to form a control problem, such a process is of little value since we could only speculate upon the existence of $O(\Pi)$.

Now, for practical system engineering, we wish to construct as well as solve control problems for sequential machines. That is, we wish to assign decision processes to sequential machines for which there exist algorithms to obtain all optimal policies. Then the three following subclasses, also given by Ibaraki [9] are to be considered. These subclasses are respectively denoted r-ℓmsdp, r-smsdp, and r-pmsdp.

Let $\Pi\ (M,h,c_o)$ be a recursive monotone sequential decision process. Π is loop-free (r,ℓmsdp) if the set of feasible policies is a finite set. Π is strictly monotone (r-smsdp) if h of Π satisfies $c < c' \Rightarrow h(c,q,x) < h(c',q,x)$ for every (c,q,x) in L_π. Similarly Π is positively monotone (r-pmsdp) if $h(c,q,x) \geq c$ for every $(c,q,x) \in L_\pi$.

The use of these classes of processes as cost functionals for the control of sequential machines will be touched upon below.

V. THE ASSIGNMENT PROBLEM

Papers by Karp and Held and by Ibaraki [10, 13], considered the problem of representing a discrete optimization problem given in the form of a discrete decision process by a finite state model, that is, a sequential decision process. In the context of the control problem for sequential machines we start with a different foundation. Essentially we have a sequential machine; our task is to characterize those sequential decision processes which we can associated with our machine to form a control problem. Then the solution of the control problem may be addressed.

In the control problem for sequential machines, we have given M, a sequential machine where $M = (Q,X,Y,\delta,\lambda)$ with an initial state q_o and a set of final states $F \subseteq Q$. To set up a control problem for the machine, as given by Definition 4, we must assign a cost function h, where

$$h : C \times Q \times X \to C$$

as well as give an initial cost c_o for starting in initial state q_o.

We can assign to any finite sequential machine, cost functions such that we may have a monotone sequential decision process. This implies that the sets accepted by finite sequential machines; (i.e. regular sets) are included in the set of feasible policies of recursive monotone sequential decision processes.

In general, the set of optimal policies for a recursive sequential decision process is included in the class of recursive sets. If every state of the machine M is a final state and there exists an optimal policy, then there exists a procedure to determine that policy. On the other hand, there is no algorithm to decide whether an arbitrarily given recursive set belongs to the set of optimal policies for some recursive-monotone sequential decision process. Likewise, there exists no algorithm to obtain an optimal policy if $O(\Pi) \neq \phi$ for an arbitrary recursive-monotone sequential decision process. These results are formally given by Ibaraki [9]. Thus we must limit ourselves at least to processes where: if there exists an optimal policy it can be determined, or the existence of the optimal policy can be determined as well as the policy itself.

The first subclass, presented in the previous section is the invertible recursive-monotone sequential decision process, denoted r-imsdp. Such a class may be considered the most general decision process to be considered as a control problem for sequential machines. Although the process of determining an $O(\Pi)$ for an r-imsdp is guaranteed to terminate, an upper bound on the number of steps cannot be determined. If a bound could be determined, then an algorithm could be given to decide whether or not an optimal policy exists. In general, if we are given a r-imsdp on a sequential machine as its control problem, and we are told $O(\Pi) \neq \phi$, then there exists a Turing machine which will always halt and will compute an optimal policy. Otherwise little more can be said about r-imsdps. If we are given only a machine M and must assign a control problem as a r-msdp, then one of the following should be utilized.

The remaining three subclasses of recursive monotone sequential decision processes have simple structures and optimal policies can be determined as well as their existence by relatively simple algorithms. In all three cases the optimal policies if they exist may be described by finite-state languages or regular sets. The simplest subclass is that of the loop-free recursive monotone sequential decision process (r-ℓmsdp). The "so-called" multistage decision processes [3] are typical examples of r-ℓmsdps. The recursive strictly monotone sequential decision process (r-smsdp) and the recursive positively monotone sequential decision process (r-pmsdp) are generalizations of shortest path problems [15, 16] and shortest path problems with non-negative arc lengths [15, 17, 18] respectively. Most important decision problems are solvable for an r-ℓmsdp as multistage decision processes.

This section has addressed the assignment of control problems for sequential machines by several subclasses of recursive-monotone sequential decision processes. In the first and most general case, the invertible r-msdps, we assumed that the control problem is given apriori for the sequential machine with knowledge that there exists an optimal policy. Otherwise, we assume we have a machine M and assign a discrete decision process to it to form a control problem. For this situation, the remaining three subclasses apply. The loop-free case is directly applicable to standard methods of multistage dynamic programming, whereas the strictly monotone and positive monotone processes require a transformation on the basic machine.

A formulation of control problems for sequential machines and several of its implications have been considered. In assigning a decision process to a sequential machine to form a control problem, we must restrict consideration to solvable classes of processes. These were considered both for the general case where the problem is given apriori as an invertible r-msdp with the existence of optimal policies may be determined. For the latter three classes, it is possible to specify the control problem as one for which dynamic programming may be applied and algorithms are referenced.

1. Sage, A.P. and J.L. Melsa, System Identification, Academic Press, New York, 1971.

2. Hopcroft, J.E. and J.D. Ullman, Formal Languages and their Relation To Automata, Addison Wesley, Mass., 1969.

3. Bellman, R.E., Dynamic Programming, Princeton University Press, Princeton, N.J., 1957.

4. Mitten, L.G., "Composition Principles for Synthesis of Optimal Multi Stage Processes", Operations Research, Vol. 12, 1964, pp 610-619.

5. Nemhanser, G.L., Introduction to Dynamic Programming, John Wiley, New York, 1966.

6. Denardo, E.V. and L.G. Mitten, "Elements of Sequential Decision Processes", J. Industrial Engineering, Vol. 18, 1967, pp 106-112.

7. Karp, R.M. and M. Held, "Finite-State Representations of Discrete Optimization Problems", SIAM J. on Computers, Vol. 2, No. 3, 1973, pp 193-210.

8. Ibaraki, T., "Finite-State Representations of Discrete Optimization Problems", SIAM J. on Computers, Vol. 2, No. 3, 1973, pp 193-210.

9. Ibid, "Solvable Classes of Discrete Dynamic Programming", J. Math. Anal. & App., Vol. 43, No. 3, 1973, pp 642-693.

10. Ibid, "Minimal Representations of Some Classes of Dynamic Programming", Working Paper, Dept. of Appld. Math & Physics, Kyoto University, Kyoto, Japan, 1973.

11. Arbib, M.A., "Tolerance Automata", Kybernetik, Vol. 3, 1967, pp 223-233.

12. Kalman, R.E., P.L. Falb and M.A. Arbib, Topics in Mathematical System Theory, McGraw Hill, New York, 1969.

13. Ibaraki, T., "Representation Theorems for Equivalent Optimization Problems", Information and Control, Vol. 21, 1973, pp 397-435.

14. Ibid, "Classes of Discrete Optimization Problems and their Decision Problems", J. Comput. & Syst. Sci., Vol. 8, 1974, pp 84-116.

15. Ford, L.R. Jr. and D. R. Fulkerson, Flows in Networks, Princeton University Press, Princeton, N.J., 1962.

16. Dreyfus, S.E., "An Appraisal of Some Shortest Path Algorithms", Operations Research, Vol. 17, 1969, pp 365-412.

17. Whiting, P.D. and J.A. Hillier, "A Method for Finding the Shortest Route Through a Road Network", Operations Research Quart., Vol. 11, 1960, pp 37-40.

18. Yen, Y.J., "Finding the Lengths of all Shortest Paths in N-Node Nonnegative Distance Complete Networks Using (1/2) N^3 Additions and N^3 Comparisons", J. Assoc. Comput. Mach., Vol. 19, 1972, pp 423-424.

A NEW THEORETICAL FRAMEWORK FOR FINITE HORIZON STOCHASTIC CONTROL[*]

STEVE SHREVE and DIMITRI BERTSEKAS
Coordinated Science Laboratory
University of Illinois
Urbana, Illinois 61801

ABSTRACT

This paper analyzes a stochastic optimal control problem over a finite horizon. The measure theoretic questions arising from an uncountable disturbance space are resolved so that the dynamic programming algorithm can be defined and employed to construct optimal and nearly optimal policies.

I. THE FINITE HORIZON STOCHASTIC OPTIMAL CONTROL PROBLEM

We take as our stochastic control model the following eight objects. The letters x and u denote elements of S and C respectively.

S - State space.

C - Control space.

Γ - Constraint set. A subset of $S \times C$. For each x, $\Gamma_x = \{u:(x,u)\epsilon\Gamma\} \neq \emptyset$.

$(W,\mathcal{F})$ - Disturbance space. A measurable space.

$p(dw|x,u)$ - Disturbance kernel. For fixed (x,u) $p(dw|x,u)$ is a probability measure on $(W,\mathcal{F})$.

f - System function. Maps SCW into S.

g - One-stage cost function. Maps SCW into $[-\infty,+\infty]$.

N - Horizon. A positive integer.

Roughly speaking the system moves from state x_k to state x_{k+1} via the system equation

$$x_{k+1} = f(x_k,u_k,w_k) , \quad k = 0,\ldots, N-1 , \tag{1}$$

and incurs cost at each stage of $g(x_k,u_k,w_k)$. The controls u_k are constrained to lie in Γ_{x_k}, $k = 0,\ldots, N-1$. The disturbances w_k are random objects with probability distributions $p(dw_k|x_k,u_k)$.

A policy $\pi = (\mu_o,\ldots,\mu_{N-1})$ is a sequence of functions $\mu_k : S \rightarrow C$ such that $\mu_k(x_k)\epsilon\Gamma_{x_k}$ for every x_k and k. When such a policy π is

---

[*] Research supported by Grant NSF ENG 74-19332.

employed, the k-th control is given by

$$u_k = \mu_k(x_k) , \quad k=0,\ldots,N-1, \tag{2}$$

where x_k is the k-th state. The cost function associated with π is

$$J_\pi(x_o) = E\{\sum_{k=0}^{N-1} g(x_k,u_k,w_k)\}. \tag{3}$$

where (1) and (2) hold. Thus defined, J_π is a function of the initial state.

The optimal cost function is

$$J^*(x_o) = \inf_\pi J_\pi(x_o), \tag{4}$$

where the infimum is over all policies π. For $\epsilon > 0$ we say a policy π is ϵ-optimal if for every $x_o \in S$

$$\begin{aligned} J_\pi(x_o) &\le J^*(x_o)+\epsilon && \text{if } J^*(x_o) > -\infty; \\ &\le -1/\epsilon && \text{if } J^*(x_o) = -\infty. \end{aligned} \tag{5}$$

A policy π is optimal if for every $x_o \in S$

$$J_\pi(x_o) = J^*(x_o). \tag{6}$$

The standard method of computing J^* and constructing an optimal or ϵ-optimal policy is to employ the dynamic programming algorithm

$$J_N(x_N) = 0 \quad \forall x_N \in S, \tag{7}$$

$$J_k(x_k) = \inf_{u_k \in \Gamma_{x_k}} E\{g(x_k,u_k,w_k) + J_{k+1}[f(x_k,u_k,w_k)]\}, \quad k=N-1,\ldots,0. \tag{8}$$

The expectation in (8) is with respect to the measure $p(dw_k|x_k,u_k)$. It is reasonable to believe that $J_o(x_o) = J^*(x_o)$ for every $x_o \in S$. If the infimum in (8) is achieved for every x_k and k, we can construct a policy $\pi = (\mu_o,\ldots,\mu_{N-1})$ such that $\mu_k(x_k)$ achieves the infimum and expect π to be optimal. Otherwise, we can choose $\mu_k(x_k)$ to nearly achieve the infimum and expect that by doing this properly, π can be constructed to be ϵ-optimal.

The difficulties with this approach center around the mathematics of expected values. The expectation operator occurs in (3) and (8) above and each time causes a problem.

Since the expectation operator can be applied only to random variables, it is necessary to impose conditions on the model which guarantee that when (1) and (2) are successively substituted in (3), the expression in braces becomes a measurable function on some probability space. Thus $p(dw|x,u)$ must have some measurability properties in (x,u), $f(x,u,w)$ and $g(x,u,w)$ must be measurable in (x,u,w), and $\mu_o,\ldots,\mu_{N-1}$ must be measurable functions from S to C. It is also necessary to impose conditions to assure that the cost-to-go functions $J_o,\dots,J_N$ defined by (7) and (8) are measurable so that the expectation in (8) is guaranteed to exist. Examples exist to show that if a function which is jointly measurable in two variables is infimized over one of the variables in the manner of (8), the result may be a nonmeasurable function.

Thus the definition of the stochastic control model given above is inadequate for the development of a general theory. We wish therefore to modify the model in such a way that:

(a) The dynamic programming algorithm (7) and (8) is defined.

(b) The cost corresponding to a policy (3) is defined.

(c) If the infimum in (8) is achieved for every x_k and k, then an optimal policy exists. Otherwise for each $\epsilon > 0$ an ϵ-optimal policy exists.

II. REVIEW OF PREVIOUS WORK

One easy way to resolve the difficulties outlined above is to assume the disturbance space W is countable, since in this case the expectation operators in (3) and (8) reduce to summations and there are no measurability considerations. The theory of dynamic programming is well developed for both a finite and an infinite horizon under this assumption ([1],[2],[3],[4]). However, by allowing only discrete disturbances, many interesting disturbance distributions are ruled out, most notably the Gaussian.

Another approach is to assume S and C are compact metric spaces, Γ is closed, $p(dw|x,u)$ is independent of (x,u), f is continuous, and g is lower semicontinuous and bounded below. Then the cost-to-go functions of (8) are lower semicontinuous, the infimum in (8) is achieved for every x_k and k, and suitably measurable functions μ_k which achieve this infimum exist ([5],[6]). The restrictiveness of these assumptions is apparent. They do, however, provide easily verifiable conditions under which an optimal policy exists and previous results in this area are extended in Theorem 3.

It is also possible to define the concept of p-essential infimum of a set of extended real-valued functions on a measure space ([7]). If S and C are measurable spaces and $p_o,\dots,p_{N-1}$ are probability measures on S, then (8) can be replaced by

$$J_k(x_k) = p_k\text{-essential } \underset{\mu_k}{\text{infimum}}\ E\{g(x_k,\mu_k(x_k),w_k) + J_{k+1}[f(x_k,\mu_k(x_k),w_k)]\}, \quad k = N-1,\dots,0, \tag{9}$$

where μ_k ranges over the set of measurable mappings from S to C which satisfy $\mu_k(x_k)\epsilon\Gamma_{x_k}$ for every x_k and k. Then J_k is measurable, $k = N-1,\dots,0$, and if the measures $p_o,\dots,p_{N-1}$ are properly chosen and the so-called countable ϵ-lattice property ([8]) is satisfied, this modified dynamic programming algorithm can be used to construct a policy π such that for p_o-almost all x_o, (5) or (6) holds, depending on whether the p_k-essential infimum in (9) is achieved for every k. We say such a π is

p_o-ϵ-optimal or p_o-optimal respectively. The major disdavantages of this approach are the need to verify the countable ϵ-lattice property and to determine $p_o,\dots,p_{N-1}$ before the dynamic programming algorithm begins. The first is equivalent to showing the existence of a p_o-ϵ-optimal policy for every p_o and the second is at least as difficult as executing the dynamic programming algorithm.

III. BOREL SPACE FRAMEWORK

The approach we deem to have the most merit is the imposition of Borel space structure on the model ([9],[10],[11],[12]). A Borel space is any nonempty Borel subset* of a complete separable metric space. We will denote by $\mathcal{B}_X$ the Borel σ-algebra in a Borel space X. An analytic subset of X is any set which can be obtained by projecting a Borel set in $X \times [0,1]$ onto X ([13],[14]) and the analytic σ-algebra $\mathcal{A}_X$ is generated by the analytic subsets of X. A universal subset E of X is any set such that for every probability measure p on $(X,\mathcal{B}_X)$, E is in the completion of $\mathcal{B}_X$ with respect to p. The σ-algebra of universal subsets of X will be denoted $\mathcal{U}_X$. Clearly $\mathcal{B}_X \subset \mathcal{A}_X$. It also holds that $\mathcal{A}_X \subset \mathcal{U}_X$ ([13]).

We assume for the remainder of the paper that S, C, and W are Borel spaces, $\mathcal{F} = \mathcal{B}_W$, Γ is analytic in $S \times C$, f and g are Borel measurable, and for fixed $B \epsilon \mathcal{B}_W$, $p(B|x,u)$ is Borel measurable in (x,u). It is possible then to show that the functions J_k, $k = 0,\dots,N-1$, are analytically measurable and that the dynamic programming algorithm can be defined by (7) and (8). We turn now to objectives (b) and (c) of Section I.

Objective (b) is met if we allow either Borel, analytically or universally measurable policies. The main result reported here is that if universally measurable policies are permitted, then objective (c) is met as well. This need not be true if policies are restricted to be Borel measurable. The key to our analysis is an extension of a selection theorem by L. D. Brown and R. Purves ([15], Theorem 2). The previous results in this area are summarized below.

Strauch [10] showed that if policies are required to be of the form $\pi = (\mu_o,\dots,\mu_{N-1})$, where for each k, μ_k is a Borel measurable function from SS to C satisfying $\mu_k(x_o,x_k) \epsilon \Gamma_{x_k}$ for every (x_o,x_k), then given a probability measure p on (S,B_S) and $\epsilon > 0$, a p-ϵ-optimal policy exists. He showed by example that an ϵ-optimal policy need not exist.

Blackwell, Freedman and Orkin ([12]) showed that if policies are of the form $\pi = (\mu_o,\dots,\mu_{N-1})$, where for each k, μ_k is an analytically measurable function from S...S to C satisfying $\mu_k(x_o,\dots,x_k) \epsilon \Gamma_{x_k}$ for every $(x_o,\dots,x_k)$, then for $\epsilon > 0$, an (everywhere) ϵ-optimal policy exists. It is not known whether an optimal policy is guaranteed to exist if the infinum in (8) is achieved for every x_k and k.

---

* A Borel subset of a topological space is a member of the smallest σ-algebra containing all open sets.

IV. MAIN RESULTS

Our model is as outlined in Section I with the Borel framework adopted in Section III. We extend the concept of policy. A policy $\pi = (\mu_o,\dots,\mu_{N-1})$ is a sequence of objects $\mu_k(du_k|x_o,u_o,\dots,u_{k-1},x_k)$ such that for each fixed $(x_o,u_o,\dots,u_{k-1},x_k)$, $\mu_k(du_k|x_o,u_o,\dots,u_{k-1},x_k)$ is a probability measure on $(C,\mathcal{B}_C)$ satisfying $\mu_k(\Gamma_{x_k}|x_o,u_o,\dots,u_{k-1},x_k)=1$, while for each fixed Borel set $B \subset C$, $\mu_k(B|x_o,u_o,\dots,u_{k-1},x_k)$ is universally measurable in $(x_o,u_o,\dots,u_{k-1},x_k)$. Thus we allow randomized policies which depend on the entire history of the system. If for each k, μ_k depends only on the initial state x_o and the current state x_k, i.e. has the form $\mu_k(du_k|x_o,x_k)$, then π is semi-Markov. If for each k, μ_k depends only on the current state x_k, π is Markov. If for each k and $(x_o,u_o,\dots,u_{k-1},x_k)$, $\mu_k(du_k|x_o,u_o,\dots,u_{k-1},x_k)$ assigns mass one to some point in Γ_{x_k}, then π is nonrandomized. By an abuse of notation, a nonradomized Markov policy can be identified with a sequence $\pi = (\mu_o,\dots,\mu_{N-1})$ of universally measurable mappings from S to C which take states x to controls in Γ_x. We will use the two representations of nonrandomized Markov policies interchangeably. From a practical point of view, these are the most desirable policies.

Let $\pi = (\mu_o,\dots,\mu_{N-1})$ be a policy. For $x_o \in S$ there is a sequence $(q_o,\dots,q_{N-1})$ of probability measures on $(S\times C, \mathcal{B}_{S\times C})$ defined by

$$q_k(B) = \int_C \int_S \cdots \int_S \mu_k(B_{x_k}|x_o,u_o,\dots,u_{k-1},x_k)$$

$$t(dx_k|x_{k-1},u_{k-1}) \cdots t(dx_1|x_o,u_o)\mu_o(du_o|x_o), \quad B\in\mathcal{B}_{S\times C},$$

where

$$t(E|x,u) = p(\{w : f(x,u,w)\in E\}|x,u), \quad E\in\mathcal{B}_S .$$

We say $(q_o,\dots,q_{N-1})$ is generated from x_o by π.

We assume one of the following conditions holds:

(F^+) For every policy π and every $x_o \in S$

$$\int_{S\times C} \int_W g(x_k,u_k,w_k)p(dw_k|x_k,u_k)q_k(d(x_k,u_k)) > -\infty, \quad k=0,\dots,N-1,$$

where $(q_o,\dots,q_{N-1})$ is generated from x_o by π.

(F^-) For every policy π and every $x_o \in S$

$$\int_{S\times C} \int_W g(x_k,u_k,w_k)p(dw_k|x_k,u_k)q_k(d(x_k,u_k)) < +\infty, \quad k=0,\dots,N-1,$$

where $(q_o,\dots,q_{N-1})$ is generated from x_o by π .

It is necessary to make one of these assumptions to assure that the indeterminate form $+\infty - \infty$ does not occur in (10) below. The symbols (F^+) and (F^-) will appear in the statements of theorems to indicate that the result holds under the appropriate assumption.

We are now in a position to give meaning to (3). For $x_o \epsilon S$, let $(q_o,\ldots,q_{N-1})$ be generated from x_o by π. Then

$$J_\pi(x_o) = \sum_{k=0}^{N-1} \int_{S\times C} \int_W g(x_k,u_k,w_k)p(dw_k|x_k,u_k)q_k(d(x_k,u_k)) \; . \tag{10}$$

Under (F^+), J_π can be shown to be a universally measurable function from S to $(-\infty,+\infty]$. Under (F^-), J_π is universally measurable from S to $[-\infty,+\infty)$. Under either (F^+) or (F^-), J^* defined by (4) can be shown to be an analytically measurable function from S to $[-\infty,+\infty]$.

We have the following theorems.

Theorem 1 (F^+) (F^-)

The dynamic programming algorithm of (7) and (8) is defined, the functions J_k, $k=0,\ldots,N-1$, are analytically measurable from S to $[-\infty,+\infty]$, and $J_o = J^*$.

Theorem 2 (F^+) (F^-)

If the infinum in (8) is achieved for every x_k and k, then there exists a nonrandomized Markov optimal policy $\pi = (\mu_o,\ldots,\mu_{N-1})$ such that $\mu_k(x_k)$ achieves this infimum for every x_k and k.

The next theorem gives continuity and compactness conditions under which the existence of a nonrandomized Markov optimal policy can be assured. It is actually a corollary of a stronger, more technical theorem not given here. Condition (i) is a generalization of a positive definite quadratic cost function in finite dimensional Euclidean space. Condition (iii) guarantees that (F^+) is in force.

Theorem 3

Assume

(i) $\Gamma = \bigcup_{j=1}^{\infty}\Gamma_j$, $\Gamma_1 \subset \Gamma_2 \subset \cdots$, Γ_j is compact $\forall j$, and

$$\lim_{j\to\infty} \inf_{(x,u)\epsilon\Gamma_j-\Gamma_{j-1}} g(x,u) = \infty;$$

(ii) f: SCW $\to$ S is continuous;

(iii) g: SCW $\to [-b, +\infty]$, where b is finite, is lower semicontinuous;

(iv) $W = R^n$;

(v) $p(dw|x,u)$ has a density $d(w;x,u)$ such that $(x_k,u_k) \to (x,u)$ implies $d(w;x_k,u_k) \to d(w;x,u)$ for Lebesgue almost every w.

Then the infimum in (8) is achieved for every x_k and k and the nonrandomized Markov optimal policy whose existence is guaranteed by Theorem 2 can be chosen to have Borel measurable components.

In general an optimal policy need not exist. We have, however, the following theorem on the existence of ϵ-optimal policies.

Theorem 4

Let $\epsilon > 0$ be given.

(F^+) There exists an ϵ-optimal nonrandomized Markov policy.

(F^-) There exists an ϵ-optimal nonrandomized semi-Markov policy and an ϵ-optimal (randomized) Markov policy.

To see that Theorem 4 under (F^-) cannot be strengthened, let $S = \{0,1,2,\ldots\}$, $C = \{1,2,\ldots\}$, $W = \{w_1,w_2\}$, $\Gamma = SC$, $N = 2$ and define

$$\begin{aligned} g(x,u) &= -u \quad \text{if} \quad x = 1 \\ &= 0 \quad \text{if} \quad x \neq 1 \end{aligned}$$

$$\begin{aligned} f(x,u,w) &= 0 \quad \text{if} \quad x = 0, \text{ or } x = 1 \text{ or } w = w_1 \\ &= 1 \quad \text{if} \quad x \neq 0,\ x \neq 1 \text{ and } w = w_2 \end{aligned}$$

$$\begin{aligned} p(\{w_1\}|x,u) &= 1 - 1/x \quad \text{if} \quad x \neq 0,\ x \neq 1 \\ p(\{w_2\}|x,u) &= 1/x \quad \text{if} \quad x \neq 0,\ x \neq 1. \end{aligned}$$

Let $\pi = (\mu_o,\mu_1)$ be a nonrandomized Markov policy. If the initial state x_o is neither zero nor one, then regardless of the policy employed, $x_1 = 0$ with probability $1 - 1/x_o$ and $x_1 = 1$ with probability $1/x_o$. Once the system reaches zero, it remains there at no further cost. If the system reaches one, it moves to $x_2 = 0$ at a cost of $-\mu_1(1)$. Thus

$$J_\pi(x_o) = -\mu_1(1)/x_o, \quad x_o \neq 0,\ x_o \neq 1$$

and

$$J^*(x_o) = -\infty, \quad x_o \neq 0,\ x_o \neq 1.$$

For any $\epsilon > 0$, π cannot be ϵ-optimal.

We note that the model considered here is stationary, i.e. S, C, Γ, W, $p(dw|x,u)$, f, and g are independent of the stage k. It is also one of perfect state information. The nonstationary model can be reduced to the stationary one and the obvious counterparts of the results given here hold in the more general case. By means of a sufficient statistic the imperfect state information model can also be reduced to the one considered here. The details appear in the first author's doctoral dissertation [16].

REFERENCES

[1] Bertsekas, D. P., Dynamic Programming and Stochastic Control, Academic Press, New York, 1976.

[2] Aoki, M., Optimization of Stochastic Systems - Topics in Discrete Time Systems, Academic Press, New York, 1967.

[3] Bellman, R., Dynamic Programming, Princeton University Press, Princeton, New Jersey, 1957.

[4] Howard, R., Dynamic Probabilistic Systems, Vols. I and II, Wiley, New York, 1971.

[5] Maitra, A., "Discounted Dynamic Programming on Compact Metric Spaces," Sankhya, Vol. 30A, pp. 211-216, 1968.

[6] Schäl, M., "Conditions for Optimality in Dynamic Programming and for the Limit of n-Stage Optimal Policies to be Optimal," Z. Wahrscheinlichkeitstheorie verw. Gebiete, Vol. 32, pp. 179-196, 1975.

[7] Dunford, N. and Schwartz, J. T., Linear Operators, Part I: General Theory, Interscience Publishers, New York, 1957.

[8] Striebel, C., Optimal Control of Discrete Time Stochastic Systems, Springer, Heidelberg, 1975.

[9] Blackwell, D., "Discounted Dynamic Programming," Ann. Math. Stat., Vol. 36, pp. 226-235, 1965.

[10] Strauch, R. C., "Negative Dynamic Programming," Ann. Math. Stat., Vol. 37, pp. 871-890, 1966.

[11] Hinderer, K., Foundations of Nonstationary Dynamic Programming with Discrete Time Parameter, Springer, Heidelberg, 1970.

[12] Blackwell, D., Freedman, D., and Orkin, M., "The Optimal Reward Operator in Dynamic Programming," Ann. Prob., Vol. 2, No. 5, pp. 926-941, 1974.

[13] Kuratowski, K., Topology I, Academic Press, New York, 1966.

[14] Parthasarathy, K., Probability Measures on Metric Spaces, Academic Press, New York, 1967.

[15] Brown, L. D. and Purves, R., "Measurable Selections of Extrema," Ann. Stat., Vol. 1, No. 5, pp. 902-912, 1973.

[16] Shreve, S., "Dynamic Programming in Complete Separable Metric Spaces," Ph.D. thesis, Dept. of Mathematics, University of Illinois, 1976.

A REACHABILITY THEORY FOR A CLASS OF UNCERTAIN DYNAMICAL SYSTEMS

B. ROSS BARMISH
Department of Engineering and Applied Science
Yale University
New Haven, Connecticut 06520

Abstract. We consider a dynamical system (S) whose impulse matrix $g(t,s)$ is known only within bounds. Within this framework, a moving target set $Y_T(t) \subseteq R^r$ is said to be strongly reachable from (S) if an admissible input $u_o(t)$ can be found steering the system output to the target — subject to worst-possible variations of $g(t,s)$ within the given bounds. The main result of this paper tells us how to convert the strong reachability problem, a problem in the infinite-dimensional input space U, into a finite-dimensional parameter optimization problem.

1. INTRODUCTION AND FORMULATION

In this paper, we shall consider a class of target set reachability problems in the presence of "model uncertainty." A dynamical system (S) will have output $y(t)$, at time $t \geq 0$, given by

$$y(t) = y_o(t) + \int_o^t g(t,s)u(s)ds \qquad \text{(S)}$$

where $y_o(t) \varepsilon R^r$ is the free response of (S), $g(t,s)$ is the impulse response matrix and $u(\cdot)$ is the input — an m-dimensional measurable function on $[0,\infty)$.

The so-called model uncertainty arises because we assume that $g(t,s)$ and $y_o(t)$ are unknown except for the fact that they are assumed to belong to some prescribed classes of functions, say G and Y_o respectively. Then, a moving target $Y_T(\cdot) \subseteq R^r$ is said to be strongly reachable (s.r.) from (S) if an admissible input $u(\cdot)$ can be found such that for some $t > 0$, we have

$$y(t) \varepsilon Y_T(t) \quad \text{for } \underline{\text{all}} \quad g \varepsilon G,\ y_o \varepsilon Y_o.$$

Various necessary and (or) sufficient conditions for strong reachability have been derived subject to "additional hypotheses" about the classes G and Y_o. For example, in [1] and [2], G and Y_o are described in terms of an input-dependent error bound between $y(t)$ and $y_m(t)$, the output of an approximate model (M) of (S). We also mention a class of strong reachability problems (e.g. see [3]-[5]) for systems corrupted by additive noise. In these papers as well, the objective is to transfer the system output (or state) to the target subject to worst-possible noise variations within given bounds.

Here, the classes G and Y_o are generated from known upper and lower bounds on the entries of $g(t,s)$ and $y_o(t)$. i.e., $g_{ij}(\cdot,\cdot)$ are measurable functions satisfying

$$g_{ij}(t,s) \varepsilon [\underline{g}_{ij}(t,s), \bar{g}_{ij}(t,s)]; \quad i = 1,2,\ldots,r; \quad j = 1,2,\ldots,m;$$

$$y_o^i(t) \varepsilon [\underline{y}_o^i(t), \bar{y}_o^i(t)]; \quad i = 1,2,\ldots,r. \qquad (*)$$

The given bounds $\underline{g}_{ij}$, $\bar{g}_{ij}, \underline{y}_o^i, \bar{y}_o^i$ are assumed to vary continuously in their arguments. Hence, the class G consists of all $r \times m$ matrices whose entries are measurable functions satisfying (*). The class Y_o is described in a similar manner.

Finally, the *target set* $Y_T(\cdot)$ at time $t \geq 0$ will be given by

$$Y_T(t) = \{y \varepsilon R^r : \underline{y}_T^i(t) \leq y^i \leq \bar{y}_T^i(t);\ \ i = 1,2,\ldots,r\}$$

i.e., a desired "range of values" is specified a priori for $y(t)$. For now, we'll consider the strong reachability problem *without* input constraints. i.e., the *admissible input set* U on $[0,t]$ will consist of all m-dimensional integrable measurable functions.

Henceforth, our objective is to develop necessary and sufficient conditions for strong reachability. Furthermore, we want our "test for strong reachability" to be finite-dimensional in nature. Such a "parameterization will be described in Theorem 4.1.

2. PRELIMINARY TRANSFORMATION

To begin, we shall reformulate the strong reachability problem as a problem of minimax over the infinite-dimensional input space U. A similar transformation is used in [4] in the context of embedding of convex sets.

2.1 *Preliminary Notation*

$$g_{ij}^o(t,s) = \tfrac{1}{2}(\underline{g}_{ij}(t,s) + \bar{g}_{ij}(t,s));$$

$$r_{ij}(t,s) = \tfrac{1}{2}(\bar{g}_{ij}(t,s) - \underline{g}_{ij}(t,s));\quad i = 1,2,\ldots,r;\quad j = 1,2,\ldots,m;$$

$$\bar{x}_m^i(t) = \bar{y}_T^i(t) - \bar{y}_o^i(t);$$

$$\underline{x}_m^i(t) = \underline{y}_T^i(t) - \underline{y}_o^i(t);\quad i = 1,2,\ldots,r.$$

2.2 *Upper and Lower Functionals*: For each fixed $t \geq 0$, we can define the following functionals on U:

$$\bar{F}_i(u,t) = \sum_{j=1}^m \int_o^t g_{ij}^o(t,s)u^j(s) + r_{ij}(t,s)|u^j(s)|ds - \bar{x}_m^i(t);$$

$$\underline{F}_i(u,t) = \sum_{j=1}^m \int_o^t -g_{ij}^o(t,s)u^j(s) + r_{ij}(t,s)|u^j(s)|ds + \underline{x}_m^i(t);\quad i=1,2,\ldots r;$$

$$F_i(t,t) = \max\{\underline{F}_i(u,t),\ \bar{F}_i(u,t)\};\quad i = 1,2,\ldots.r;$$

$$F(u,t) = \max_{i \leq r} F_i(u,t).$$

2.2.1. *Remark*: It is easily shown that all four functionals above are convex with respect to u. We shall exploit this fact in Section 3.

2.3 *Lemma* (Reformulated Problem); $Y_T(\cdot)$ is s.r. *from* (S) *if and only if* $\exists t \geq 0$ *such that*

$$\min_{u \varepsilon U} F(u,t) \leq 0. \qquad (\mathbb{P})$$

Proof: We seek an input $u \varepsilon U$ and a time $t > 0$ such that the maximum and minimum excursions of $y^i(t)$ are contained within $[\underline{y}^i_T(t), \bar{y}^i_T(t)]$ (as g_{ij} and y^i_o roam over the sets prescribed in (*)). Analytically, we require

$$\bar{y}^i_o(t) + \sum_{j=1}^{m} \sup_{g_{ij}}\{\int_o^t g_{ij}(t,s)u^j(s)\,ds\} \le \bar{y}^i_T(t); \tag{2.3.1}$$

$$\underline{y}^i_o(t) + \sum_{j=1}^{m} \inf_{g_{ij}}\{\int_o^t g_{ij}(t,s)u^j(s)\,ds\} \ge \underline{y}^i_T(t); \quad i = 1,2,\ldots r. \tag{2.3.2}$$

Now,

$$\sup_{g_{ij}}\{\int_o^t g_{ij}(t,s)u^j(s)\,ds\} = \int_o^t g^o_{ij}(t,s)u^j(s)\,ds + \sup_{|\eta_{ij}(t,\cdot)| \le r_{ij}(t,s)} \int_o^t \eta_{ij}(t,s)u^j(s)\,ds.$$

It is then easily shown that

$$\eta_{ij}(t,s) = r_{ij}(t,s)\,\mathrm{sgn}\, u^j(s)$$

attains the supremum above. (Note: the continuity assumption on $\underline{g}_{ij}$, $\bar{g}_{ij}$ guarantees that r_{ij} (hence η_{ij}) is measurable.) We obtain

$$\sup_{g_{ij}} \int_o^t g_{ij}(t,s)u^j(s)\,ds = \int_o^t g^o_{ij}(t,s)u^j(s)\,ds + \int_o^t r_{ij}(t,s)|u^j(s)|\,ds. \tag{2.3.3}$$

Substitution of (2.3.3) into (2.3.1), yields the condition $\bar{F}_i(u,t) \le 0$ for $i = 1,2,\ldots r$. Similarly, the infimum condition in (2.3.2) can be shown equivalent to $\underline{F}_i(u,t) \le 0$. Consequently, $Y_T(\cdot)$ is s.r. from (S) if and only if $\exists t \ge 0$, $\exists u \varepsilon U$ such that

$$\max_i \max \{\underline{F}_i(u,t), \bar{F}_i(u,t)\} \le 0$$

This is precisely the requirement in (P). (Note: we have subsumed details about the attainment of the inf in (P)).

3. PARAMETERIZATION OF REFORMULATED PROBLEM

It is well known that many infinite-dimensional extremum problems admit a finite-parameterization under the transformation $f \to f^*$, the conjugacy correspondence of Fenchel (see [5]-[9]). If U^* denotes the collection of bounded linear functionals on U, then $F^*(\cdot,t): U^* \to R \cup \{+\infty\}$, the Fenchel conjugate of $F(\cdot,t)$ (for fixed t) is given by

$$\begin{aligned} F^*(u^*,t) &= \sup_{u \varepsilon U} [u^*(u) - F(u,t)]; \quad u^* \varepsilon U^* \\ &= \sup_{u \varepsilon U} [\int_o^t \langle u^*(s), u(s)\rangle ds - F(u,t)] \end{aligned} \tag{3.0.0}$$

The following lemma is a trivial consequence of Lemma 2.3 and the definition of F*.

3.1 *Lemma*: $Y_T(\cdot)$ *is s.r. from* (S) iff $\exists t \geq 0$ *such that* $F^*(0,t) \geq 0$.

Consequently, we hope to parameterize $F^*(0,t)$. To meet this end, we introduce (i) 2 r-dimensional parameter vectors $\underline{x}$ and $\bar{x}$, (ii) a set of 2r m-dimensional measurable functions $\underline{u}^*_1(\cdot),\underline{u}^*_2(\cdot),\ldots,\underline{u}^*_r(\cdot)$, $\bar{u}^*_1(\cdot),\ldots,\bar{u}^*_r(\cdot)$. In the following lemma, we express $F^*(0,t)$ in terms of these new variables and the known data.

3.2 *Lemma* (Partial Parametrixation):

$$F^*(0,t) = \min \{ \sum_{i=1}^{r} \bar{x}^i \bar{x}^i_m(t) - \underline{x}^i \underline{x}^i_m(t)\} \tag{3.2.0}$$

subject to

P1. $\sum_{i=1}^{r} \bar{x}^i \bar{u}^*_i + \underline{x}^i \underline{u}^*_i = 0; \; u^*_i = (u^*_{i1},u_{i2},\ldots,u^*_{im})$

P2. $\bar{u}^*_{ij}(s)\varepsilon[\underline{g}_{ij}(t,s),\bar{g}_{ij}(t,s)]; \; i = 1,2,\ldots r; \; j = 1,2,\ldots m$

P3. $-\underline{u}^*_{ij}(s)\varepsilon[\underline{g}_{ij}(t,s), \bar{g}_{ij}(t,s)]; \; i = 1,2,\ldots,r; \; j = 1,2,\ldots m,$

P4. $\bar{x}^i \geq 0, \; \underline{x}^i \geq 0; \; i = 1,2,\ldots r;$

P5. $\sum_{i=1}^{r} \bar{x}^i + \underline{x}^i = 1.$

Proof: Theorem 1.5 of [8] enables us to express $F^*(0,t)$ in terms of its constituents $\bar{F}^*_i(0,t)$, $\underline{F}^*_i(0,t)$. i.e.,

$$F^*(0,t) = \min \{ \sum_{i=1}^{r} \bar{x}^i \bar{F}^*_i(\bar{u}^*_i,t) + \underline{x}^i \underline{F}^*_i(\underline{u}^*_i,t)\} \tag{3.2.1}$$

subject to P1, P4 and P5. Now, for fixed i,

$$\bar{F}^*_i(\bar{u}^*_i,t) = \sup_{u\varepsilon U}[\bar{u}^*_i(u) - \bar{F}_i(u,t)]$$

$$= \bar{x}^i_m(t) + \sup_{u\varepsilon U}[\sum_{j=1}^{m} \int_0^t (\bar{u}^*_{ij}(s) - g^o_{ij}(t,s))u^j(s) - r_{ij}(t,s)|u^j(s)|ds]$$

$$\leq \bar{x}^i_m(t)+\sup_{u\varepsilon U} \sum_{j=1}^{m} (\int_0^t |u^j(s)|ds)\cdot\sup_{s\varepsilon[0,t]} (|\bar{u}^*_{ij}(s) - g^o_{ij}(t,s)|- r_{ij}(t,s)) \tag{3.2.2}$$

Consequently, we are led to consider 2 cases.

Case 1: $|\bar{u}^*_{ij}(s) - g^o_{ij}(t,s)| \leq r_{ij}(t,s)$ almost all $s\varepsilon[0,t]$ and all $j = 1,2,\ldots m$. From (3.2.2), it is obvious that $u^j(s) \equiv 0$ attains the supremum. Hence, $\bar{F}^*_i(\bar{u}^*_i,t) = \bar{x}^i_m(t)$.

Case 2: $|\bar{u}^*_{ij}(s) - g^o_{ij}(t,s)| < r_{ij}(t,s)$ for s belonging to some set τ of positive measure (for at least one j, say $j = k$.) Now we claim that $\bar{F}^*_i(\bar{u}^*_i,t) = \infty$. To prove this, we let

$$\delta = \int_\tau \{|\bar{u}^*_{ik}(s) - g^o_{ik}(t,s)| - r_{ik}(t,s)\}ds$$

Clearly, $\delta > 0$. We also define the sequence

$$\begin{aligned} u_n^j(s) &= n \operatorname{sgn} (\bar{u}^*_{ij}(s) - g^o_{ij}(t,s)) \text{ for } s\varepsilon\tau \text{ and } j = k; \\ &= 0 \text{ otherwise.} \end{aligned}$$

Then

$$\begin{aligned} \bar{F}^*_i(\bar{u}^*_i,t) &\geq \sup_n [\bar{u}^*_i(u_n) - \bar{F}_i(u_n,t)] \\ &= \bar{x}^i_m(t) + \sup_n n\delta \\ &= \infty . \end{aligned}$$

Hence, we obtain

$$\begin{aligned} \bar{F}^*_i(\bar{u}^*_i,t) &= \bar{x}^i_m(t) \quad \text{when } \bar{u}^*_i \text{ satisfies P2;} \\ &= \infty \quad \text{otherwise.} \end{aligned}$$

By a similar argument,

$$\begin{aligned} \underline{F}^*_i(\underline{u}^*_i,t) &= -\underline{x}^i_m(t) \text{ when } \underline{u}^*_i \text{ satisfies P3;} \\ &= \infty \text{ otherwise.} \end{aligned}$$

Substitution for $\underline{F}^*_i$, $\bar{F}^*_i$ in (3.2.1) then yields the desired result.

3.2.1 Remarks: Motivated by the fact that the infinite-dimensional variables $\underline{u}^*_i$, $\bar{u}^*_i$ do not appear in the objective function of (3.2.0), we hope to obtain a description of $F^*(0,t)$ which depends only on $\underline{x}$ and $\bar{x}$. Consequently we define

$$\Omega_t = \{(\underline{x},\bar{x})\varepsilon R^{2r}\text{: P4, P5 hold and P1, P2, P3 hold for some } \underline{u}^*_i,\bar{u}^*_i\}$$

Clearly, P1-P5 is then equivalent to the condition $(\underline{x},\bar{x})\varepsilon\Omega_t$. Now, we shall describe conditions for membership in Ω_t involving $\underline{x},\bar{x}$ but not $\underline{u}^*_i,\bar{u}^*_i$.

3.2.2. R^j and G_o: For fixed $(\underline{x},\bar{x},t)$, we define the following scalar functions of s:

$$R^j(\underline{x},\bar{x},t,s) = \sum_{i=1}^{r} r_{ij}(t,s)(\bar{x}^i + \underline{x}^i);\ j = 1,2,\ldots,m;$$

$$G^j_o(\underline{x},\bar{x},t,s) = \sum_{i=1}^{r} g^o_{ij}(t,s)(\bar{x}^i - \underline{x}^i);\ j = 1,2,\ldots,m.$$

Then we have

3.3 Lemma: $(\underline{x},\bar{x})\varepsilon\Omega_t$ if and only if P4 and P5 hold and

$$|G^j_o(\underline{x},\bar{x},t,s)| \leq R^j(\underline{x},\bar{x},t,s) \text{ for almost all } s\varepsilon[0,t];\ j = 1,2,\ldots,m. \quad (3.3.0)$$

Proof: (Sufficiency) Suppose $(\underline{x},\bar{x})$ satisfies P4, P5 and (3.3.0). Then we claim that P1-P3 will be satisfied by making the following choice for $\underline{u}^*_{ij}(s),\bar{u}^*_{ij}(s)$:

$$\bar{u}^*_{ij}(s) = g^o_{ij}(t,s) \quad \underline{\text{if}} \quad R^j(\underline{x},\bar{x},t,s) = 0;$$

$$= g^o_{ij}(t,s) - \frac{r_{ij}(t,s)}{R^j(\underline{x},\bar{x},t,s)} \cdot G^j_o(\underline{x},\bar{x},t,s) \text{ otherwise.}$$

$$\underline{u}^*_{ij}(s) = - g^o_{ij}(t,s) \quad \underline{\text{if}} \quad R^j(\underline{x},\bar{x},t,s) = 0;$$

$$= - g^o_{ij}(t,s) - \frac{r_{ij}(t,s)}{R^j(\underline{x},\bar{x},t,s)} \cdot G^j_o(\underline{x},\bar{x},t,s) \text{ otherwise.}$$

Now, P2 and P3 are trivially satisfied if $R^j(\underline{x},\bar{x},t,s) = 0$. Otherwise, we look at the quantity

$$|\bar{u}^*_{ij}(s) - g^o_{ij}(t,s)| \leq r_{ij}(t,s) \cdot \frac{|G^j_o(\underline{x},\bar{x},t,s)|}{R^j(\underline{x},\bar{x},t,s)}$$

$$\leq r_{ij}(t,s) .$$

i.e., $g^o_{ij}(t,s) - r_{ij}(t,s) \leq \bar{u}^*_{ij}(s) \leq g^o_{ij}(t,s) + r_{ij}(t,s)$

Equivalently, $\bar{u}^*_{ij}(s)$ satisfies P2. Similarly, it follows that $\underline{u}^*_{ij}(s)$ satisfies P3. To verify P1, we look at

$\sum_{i=1}^{r} \bar{x}^i \bar{u}^*_{ij}(s) + \underline{x}^i \underline{u}^*_{ij}(s) = 0$ if $R^j(\underline{x},\bar{x},t,s) = 0$. Otherwise,

$$= \sum_{i=1}^{r} \bar{x}^i \Big(g^o_{ij}(t,s) - \frac{r_{ij}(t,s)}{R^j(\underline{x},\bar{x},t,s)} \cdot G^j_o(\underline{x},\bar{x},t,s)\Big)$$

$$+ \underline{x}^i \Big(-g^o_{ij}(t,s) - \frac{r_{ij}(t,s)}{R^j(\underline{x},\bar{x},t,s)} G^j_o(\underline{x},\bar{x},t,s)\Big)$$

$$= \sum_{i=1}^{r} (\bar{x}^i - \underline{x}^i) g^o_{ij}(t,s) - \frac{G^j_o(\underline{x},\bar{x},t,s)}{R^j(\underline{x},\bar{x},t,s)} \sum_{i=1}^{r} (\bar{x}^i + \underline{x}^i) r_{ij}(t,s)$$

$= 0.$

(Necessity) We suppose that $(\underline{x},\bar{x}) \varepsilon \Omega_t$ and we must verify (3.3.0). Let $\underline{u}^*_{ij}(s)$, $\bar{u}^*_{ij}$ be appropriate time functions satisfying P1-P3. Then,

$$
\begin{aligned}
|G_o^j(\underline{x},\bar{x},t,s)| &= \Big|\sum_{i=1}^{r} \bar{x}^i(g_{ij}^o(t,s)-\bar{u}^*_{ij}(s))-\underline{x}^i(g_{ij}^o(t,s)+\underline{u}^*_{ij}(s))\Big| \text{ by P1} \\
&\le \sum_{i=1}^{r} \bar{x}^i|g_{ij}^o(t,s)-\bar{u}^*_{ij}(s)|+\underline{x}^i|g_{ij}^o(t,s)+\underline{u}^*_{ij}(s)| \\
&\le \sum_{i=1}^{r} \bar{x}^i r_{ij}(t,s)+\underline{x}^i r_{ij}(t,s) \\
&= R^j(\underline{x},\bar{x},t,s).
\end{aligned}
$$

4. MAIN RESULT

The analysis of Section 3 is summarized by the following theorem.

4.1 *Theorem:* $Y_T(\cdot)$ *is s.r. from* (S) *if and only if* $\exists t \ge 0$ *such that*

$$\min\{V(\underline{x},\bar{x},t):(\underline{x},\bar{x})\in\Omega_t\} \ge 0$$

where

$$V(\underline{x},\bar{x},t) = \sum_{i=1}^{r} \bar{x}^i\bar{x}_m^i(t) - \underline{x}^i\underline{x}_m^i(t)$$

and $\Omega_t \subseteq R^{2r}$ is the (convex) region described by

Ω1. $\bar{x}^i \ge 0$; $\underline{x}^i \ge 0$; $i = 1,2,\ldots, r$;

Ω2. $\sum_{i=1}^{r} \bar{x}^i + \underline{x}^i = 1$;

Ω3. $|G^j(\underline{x},\bar{x},t,s)| \le R^j(\underline{x},\bar{x},t,s)$ for almost all $s\in[0,t]$; $j = 1,2,\ldots,m$.

5. REFINEMENTS

The definition of strong reachability can be modified to include input constraints. Below, we provide a refinement of Theorem 4.1 for the case of "limited fuel."

5.1 *Theorem:* $Y_T(\cdot)$ *is s.r. from* (S) *with fuel constraint*

$$\int_0^{t_a} \sum_{i=1}^{m} |u^i(s)|ds \le M \quad (t_a = \text{arrival time at } Y_T) \qquad \text{(f)}$$

if and only if $\exists t \ge 0$ *such that*

$$\min\{V_f(\underline{x},\bar{x},t):(\underline{x},\bar{x})\in\Omega_f\} \ge 0$$

where Ω_f is described by Ω1 and Ω2 and

$$V_f(\underline{x},\bar{x},t) = V(\underline{x},\bar{x},t) + M \max_{j,s\in[0,t]} \Delta^j(\underline{x},\bar{x},t,s);$$

$$\Delta^j(\underline{x},\bar{x},t,s) = \max\{0, \ |G^j(\underline{x},\bar{x},t,s)|-R^j(\underline{x},\bar{x},t,s)\} .$$

6. REFERENCES

[1] V. Mauro and G. Menga, "Target Set Control of Norm Uncertain Systems," Proc. 3rd IFAC Symp. on Sensitivity, Adaptivity and Optimality, Ischia, 1973.

[2] B. R. Barmish, "Target Set Reachability Criteria for Dynamical Systems Described by Inaccurate Models," to appear.

[3] D. P. Bertsekas and I. B. Rhodes, "On the Minimax Reachability of Target Sets and Target Tubes," Automatica, vol. 7, pp. 233-237, 1971.

[4] M. C. Delfour and S. K. Mitter, "Reachability of Perturbed Systems and MinSup Problems," SIAM J. Control, vol. 7, pp. 521-533, 1969.

[5] H. S. Witsenhausen, "A Minimax Control Problem for Sampled Linear Systems," IEEE Trans. Auto. Control, AC-13, pp. 5-21, 1968.

[6] B. R. Barmish, "Robust Solutions of Perturbed Linear Equations," to appear in December 1976 issue of IEEE Trans. on Auto. Control.

[7] D. G. Luenberger, Optimization by Vector Space Methods, Wiley, New York, 1967.

[8] A. D. Ioffe and V. M. Tikhomirov, "Duality of Convex Functions and Extremum Problems," Russian Math. Surveys, vol. 23, pp. 53-124, 1968.

[9] R. T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, New Jersey, 1972.

COMPUTATIONAL METHODS TO OBTAIN TIME OPTIMAL JET ENGINE CONTROL*

R. J. BASSO AND R. J. LEAKE
Electrical Engineering Department
University of Notre Dame
Notre Dame, Indiana 46556

ABSTRACT

Dynamic Programming and the Fletcher-Reeves Conjugate Gradient Method are two existing methods which can be applied to solve a general class of unconstrained fixed time, free right end optimal control problems. New techniques are developed to adapt these methods to solve a time optimal control problem with state variable and control constraints. Specifically, they are applied to compute a time optimal control for a jet engine control problem.

1. INTRODUCTION

Consider the standard discrete time invariant system

$$x(t+\Delta t) = x(t) + f(x(t), x(t)) \qquad x(k) = x \tag{1.1}$$

with starting time k, terminal time N and a performance index function

$$J(x,\underline{u}) = Y(x(N)) + \sum_{t=k}^{N-1} L(x(t), u(t), t) \tag{1.2}$$

to be minimized. In this system, $x(t)$ is an n-dimensional state variable vector, $u(t)$ an m-dimensional control vector, and $\underline{u}$ denotes the sequence

$$\underline{u} = u(k), u(k+1), \ldots, u(N-1) \tag{1.3}$$

where each component is a member of a control set $U(x)$.

This paper extends two nonlinear global mathematical programming methods, which can solve the above problem, to solve the time optimal control problem with state variable and control constraints of the form

$$Ax(t) + Bu(t) \leq C \tag{1.4}$$

Here, the terminal time is defined as the first instant at which the system reaches a designated target state x_F, which is an equilibrium point of the nonlinear system description.

2. JET ENGINE CONTROL PROBLEM

A reduced single spool turbojet engine, derived in [1] and [2], is described by the difference equations

$$\rho_B(t+\Delta t) = \rho_B + \Delta t(37.78\dot{\omega}_3 - 38.448P_4 + .66849) \tag{2.1}$$

$$N_c(t+\Delta t) = N_c + \Delta t\left(\frac{1.258}{N_c}\right)\left(\frac{P_4^2}{\rho_B} - \dot{\omega}_3 N_c^2\right) \tag{2.2}$$

---

*This research was supported by National Aeronautics and Space Administration Grant NASA NSG 3048.

$$\dot{\omega}_3 = 1.3009N_c - .139825P_4 - .13982\sqrt{P_4^2 + .41688N_c^2 - .0899P_4N_c} \tag{2.3}$$

with constraints

$$P_4 \le 1.25\rho_B \text{ (turbine inlet temperature constraint)} \tag{2.4}$$

$$P_4 \le 1.25N_c \text{ (surge margin)} \tag{2.5}$$

where ρ_B is the burner density, N_c the rotational speed, $\dot{\omega}_3$ the compressor mass discharge flow, and P_4, the burner pressure, is the control variable. In reality, P_4 is a state variable of a higher order model but it is most heavily influenced by the fuel flow, a physically controllable variable. The control problem is to find the shortest control sequence $\underline{P_4}$ which takes the jet engine from an idle condition, or windmill,

$$x = \begin{bmatrix} \rho_B = 1.774 \\ N_c = .5461 \end{bmatrix} \tag{2.6}$$

to a high thrust state

$$x_F = \begin{bmatrix} \rho_B = 1 \\ N_c = 1 \end{bmatrix} \tag{2.7}$$

such that the surge margin and turbine inlet temperature constraints are satistifed.

3. A SUCCESSIVE APPROXIMATIONS TECHNIQUE

Dynamic Programming is a method well suited for solving low order, fixed time, free right end optimal control problems with state variable and control constraints. A successive approximations technique extends the class of problems to the free time, fixed right end optimal control problem. Let K denote the integers, X an arbitrary nonempty state set, and $S \subset X \times K$ an arbitrary nonempty target set. Suppose we also have a control constraint U(x) set which depends upon the state x and is such that controls in U(x) guarantee that the next state is in the state set X. Only the assumption that

$$V_k(x) = \min_{\underline{u}} J_k(x,\underline{u}) \tag{3.1}$$

exists is required to prove Bellman's Conditions [4], [5], and [6],

$$V_k(x) = \min_{u \in U(x)} \Big[L(x,u,k) + V_{k+1}(x+f(x,u)) \Big] \quad (x,k) \notin S \tag{3.2}$$

$$V_k(x) = Y(x) \qquad (x,k) \in S$$

are necessary and sufficient for optimality. These are the usual equations of Dynamic Programming. Note that a particular $u = v(x,k) \in U(x)$ which minimizes the above expression can be used to make up $v(x,k)$ for all $(x,k) \in S$ and comprises an optimal feedback control law.

An alternative method for solving problems of this class called "successive approximations" was suggested by Bellman [4] and later developed by Leake, Liu, and Richardson in [6] and [7].

Let $V_k{}^n(x)$ be any function such that $V_k{}^n(x) \geq V_k(x)$ and let $v^n(x,k)$ be a control law which results when performing the minimization

$$\min_{u\varepsilon U(x)} \left[L(x,u,k) + V_{k+1}{}^n(x+f(x,u)) \right] \quad (x,k) \notin S \tag{3.3}$$

It is shown in [6] that if $V_k{}^{n+1}(x)$ is the performance function resulting from $v^n(x,k)$, we have

$$V_k(x) \leq V_k{}^{n+1}(x) \leq V_k{}^n(x) \tag{3.4}$$

and further, $V_k{}^n(x)$ converges monotonically to $V_k(x)$ in a finite number of steps although each (x,k) may require a different number of steps.

In the special case (the time optimal control problem) where $V_k(x)$ is independent of k, $V_k(x) = V(x)$, the approximating sequence $V_k{}^n(x) = V^n(x)$, and the minimization becomes

$$V^{n+1}(x) = \min_{u\varepsilon U(x)} \left[L(x,u,k) + V^n(x+f(x,u)) \right] \quad (x,k) \notin S \tag{3.5}$$

We now define a standard Dynamic Programming problem which yields a successive approximation solution to the time optimal problem of interest. Let the system be the same as before with target set

$$S = S_1 \cup S_2 \tag{3.6}$$

where S is illustrated in Figure 1. and

$$S_1 = \{(x,k): \ k = 0, \ \ x \ \varepsilon \ X\}$$

$$S_2 = \{(x,k): \ k < 0, \ \ x = x_F\}$$

Let X and $U(x)$ be as above, let $N = 0$, let $L(x,u,k) = 1$, and let $Y(x)$ be chosen such that

$$\begin{aligned} Y(x) &\geq V(x) \\ Y(x_F) &= 0 \end{aligned} \tag{3.7}$$

The Bellman's Conditions are as in (3.2) and the problem can be solved as an ordinary Dynamic Programming problem, remembering that $(x_F,k) \ \varepsilon \ S$ so that

$$V_k(x_F) = 0 \tag{3.8}$$

for all $k \leq N = 0$.

Note that the actual target set of interest is simply the set of all (x,k) where $x = x_F$. We establish a successive approximation of $V(x)$ by choosing

$$V^{o}(x) = Y(x) \tag{3.9}$$

and equate

$$V_k(x) = V^{-k}(x) \qquad k = 0, -1, -2, \ldots \tag{3.10}$$

Then, Bellman's Conditions are equivalent to

$$V^{n+1} = \min_{u \varepsilon U(x)} \left[L(x,u,k) + V^{n}(x+f(x,u)) \right] \quad (x,k) \notin S \qquad n = 0, 1, 2, \ldots \tag{3.11}$$

which is simply the successive approximation equation. This establishes that the Dynamic Programming problem actually gives a successive approximations solution, with

$$V^{n}(x) \downarrow V(x) \tag{3.12}$$

and associated feedback control law

$$v^{n}(x) \rightarrow v(x) \tag{3.13}$$

In practice, the state set X is discretized and linear interpolation is used to get approximate solutions, but convergence still occurs. Choosing some k_T as time when convergence is adequate, we equate

$$V^{-k_T} = V(x) \tag{3.14}$$

where V(x) represents the minimal number of steps needed for

$$v(x,k_T) = v(x) \tag{3.15}$$

to take the system from initial state x to final state x_F.

In the actual discretization, the relative speed of the state variables must be considered and a suitable scale developed so that the motion of one time step does not exceed one increment in any state variable. Larson in [5] suggests a useful relation for this purpose

$$\Delta t = \min_{\substack{i=1,2,\ldots,n \\ j=1,2,\ldots,m}} \frac{\Delta x_i}{|F_i(x_i(t),u_j(t))|} \tag{3.16}$$

where

$$F_i(x_i(t),u_j(t)) = f_i(x_i(t),u_j(t))/\Delta t \tag{3.17}$$

The chosen discretization must include the target set and several states along the constraint boundaries.

In performing the minimization of (3.11), an interpolation scheme is often required to calculate $V^n(x+f(x,u))$. The following relation, illustrated in Figure 2., is found to best satisfy the trade-off between accuracy and computational requirements.

$$V_p = (1-a)(1-b)V_{i,j} + b(1-a)V_{i,j+1} + a(1-b)V_{i+1,j} + ab\ V_{i+1,j+1} \quad (3.18)$$

This interpolation always approximates V^n high and the approximating sequence may not converge completely in a finite number of steps, however $V^n(x)$ is found to stabilize. Recall, that before each iteration, the cost function $V^n(x_F)$ must be reset equal to zero.

4. MATHEMATICAL PROGRAMMING

The standard optimal control problem defined in the introduction can be regarded as the minimization of a function

$$J(\underline{u}) = K(x(N)) + \sum_{t=0}^{N-1} L(x(t),u(t),t) \quad (4.1)$$

of several variables. In this section, special techniques are applied to an existing Fletcher-Reeves Conjugate Gradient Algorithm [8] so that the time optimal control problem with state variable end control constraints can be solved using these concepts. This algorithm can then be directly applied to the jet engine problem with $J(\underline{u})$ being evaluated by integrating the system with a nominal control sequence for $t = 0, 1, 2, \ldots N - 1$. The gradient

$$\nabla uJ = [\nabla_{u(0)}J \quad \nabla_{u(1)}J \ldots \nabla_{u(N-1)}J] \quad (4.2)$$

is determined by solving the adjoint equations

$$\lambda(t) = \lambda(t+1) + \lambda(t+1)\nabla_x f(x(t),u(t)) + \nabla_u L(x(t),u(t),t) \quad (4.3)$$

$$\lambda(N) = \nabla_x K(x(N))$$

for $t = N - 1, \ldots, 2, 1$ and observing that

$$\nabla_{u(t)}J = \lambda(t+1)\nabla_u f(x(t),u(t)) + \nabla_u L(x(t),u(t),t) \quad (4.4)$$

The direction for descent is a linear combination of the present gradient and previous direction. After penalty methods and barrier functions were found inadequate to simulate the control constraints, a simple feasible directions idea was added to the algorithm. Whenever the new direction causes a new control component to violate constraints, the component is reset at the nearest constraint boundary; thus, the new control is completely "legal."

To produce a time optimal solution, the system is solely penalized for not being at the target and

$$L(x(t),u(t),t) = C^T(x(t)-x_F)N\left(\frac{t}{N}\right)^5 \quad (4.5)$$

where C is a vector of constants weighted in proportion to the time constants of the state variables, which are obtained by computing the eigenvalues of the Jacobian Matrix evaluated at x_F. A heavier penalty is applied to the slower reacting variables to suppress oscillatory solutions which can appear when all c components are equal.

The fact that the problem is time optimal and the optimal number of time steps N is initially unknown does not offer any real difficulty as a few rough computational runs quickly gives one a good estimate of N. The final term is included in (4.5) to provide a penalty only for later times in the trajectory; this gives the system freedom to quickly react.

A quick, accurate line search to find

$$h(\alpha^*) = \min_{\alpha \geq 0} J(\underline{u} + \alpha \underline{d}) \tag{4.6}$$

is a vital component of the modified Fletcher-Reeves method. The line search proposed here fits a cubic polynomial through three test points and with derivative information algebraically solves for α^* in one step. Assuming that $h(\alpha)$ obeys the necessary convexity requirements illustrated in Figure 3., let

$$h(\alpha) = h(0) + h'(0)\alpha + A\alpha^2 + B\alpha^3 \tag{4.7}$$

and a straightforward minimization procedure gives

$$\alpha^* = \frac{-A+\sqrt{A^2-3Bh'(0)}}{3B} \tag{4.8}$$

where

$$A = h(a) - \frac{a^3}{b^3} h(b) + h(0)\left(\frac{a^3}{b^3} - 1\right) + h'(0)\left(\frac{a^3}{b^3} - a\right) \tag{4.9}$$

$$B = \frac{h(a) - h(0) - h'(0)a + Aa^2}{a^3} \tag{4.10}$$

Since $h(0)$ and $h'(0)$ are already known, α^* is directly calculated once suitable values of a, b, h(a), and h(b), which satisfy the conditions of Fig. 3, are obtained. To quickly find these values, a subroutine which features information sensitive tests has been developed in [10] and is illustrated in Figure 4. With a maximum of three to four trajectory calculations, this subroutine is able to produce accurate α^* values which span a wide range of several orders of magnitude.

5. SOLUTIONS TO THE JET ENGINE CONTROL PROBLEM

The software which implements the two above methods is written to solve a general class of nonlinear time optimal control problems of n^{th} order and $m \leq n$ controls. A feedback control law for the jet engine was obtained from the successive approximations sequence with a step size of $\Delta t = .001$. Figure 5. illustrates the feedback control law and the resulting motion of the system due to the control law. Three regions of control are evident.

1) P_4 rides the surge margin constraint ($P_4 = 1.25N_c$)
2) P_4 switches and rides the turbine inlet temperature constraint ($P_4 = 1.25\rho_B$)
3) P_4 instantly drops to its minimum value and then jumps to the equilibrium value of 1. This is a bang-bang control.

Note that the control strategy is similar for any starting state in regions 1 and 2 and that all motions in these regions lead to a common path in region 2. In Figure 6., the time response, which results from applying the feedback control law to our problem of interest, is plotted. The optimal time is .8 seconds.

In Figure 7., a comparison is made between the open loop control solution calculated by the modified Fletcher-Reeves Conjugate Gradient Method of section 4 with a time step size of Δt - .005 and the solution of Fig. 6. Three control regions are seen but the rapid switching action of Figure 6 is not evident. However, an additional run with the initial state taken along the state variable trajectory near the end of region 2 and a smaller time step size of $\Delta t = .001$ produced the bang-bang control illustrated by the dotted graph of Figure 8. The smaller time step allowed the algorithm to detect the rapid switching action and to calculate a solution similar to that produced by the successive approximation sequence.

6. SUMMARY

The objective of this work is to adapt two well known computational methods to solve the general class of fixed right end, free time optimal control problems with state variable and control constraints. A successive approximations technique applied to a standard Dynamic Programming algorithm can solve this class of low-order optimal control problems to produce a time optimal feedback control law.

A feasible directions idea, used in conjunction with penalty functions, extends the Fletcher-Reeves Conjugate Gradient Algorithm to produce a time optimal open loop control sequence for the general class of time optimal control problems with state variable and control constraints. Computational requirements for this method increase only arithmetically with the system order.

The software which solves the jet engine problem is generalized to solve any general optimal control problem of this class.

REFERENCES

1. Seldner, Kurt; Geyser, Lucile C.; Gold, Harold; Walker, Darrell; and Burgner, Gary. Performance and Control Study of a Low-Pressure-Ratio Turbojet Engine for A Drone Aircraft, NASA TM X-2537, 1972.

2. Brennen, T. C., "Simplified Simulation Models for Control Studies of Turbojet Engines," Master's Thesis, Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana, August, 1975.

3. Pontryagin, L. S., The Mathematical Theory of Optimal Processes, New York: Interscience, 1962.

4. Bellman, R. E., Dynamic Programming, Princeton: Princeton University Press, 1957.

5. Larson, R., State Increment Dynamic Programming, New York: American Elserrier Publishing Company, 1968.

6. Richardson, M. H.; and Leake, R. J. "Optimization of Discrete-Time Systems by Successive Approximations," College of Engineering, University of Notre Dame, Notre Dame, Indiana, 1968.

7. Leake, R. J.; and Liu, Ruey-wen. "Construction of Suboptimal Control Sequences," J. SIAM Control, Vol.5 No. 1, 1967, 54-63.

8. Luenberger, D., Introduction to Linear and Nonlinear Programming, Reading Massachusetts: Addison-Wesley, 1965.

9. Lasdon, L. S., "The Conjugate Gradient Method for Optimal Control Problems," IEEE Transactions on Automatic Control, April, 1967.

10. Basso, R. J., "Computation Alternatives to Obtain Time Optimal Jet Engine Control," Master's Thesis, Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana, August, 1976.

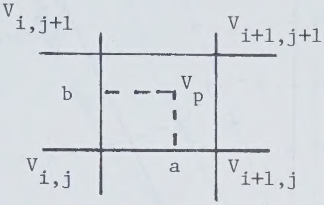

$V_{i,j+1}$
$V_{i+1,j+1}$
b
V_p
a
$V_{i,j}$
$V_{i+1,j}$

Figure 2.

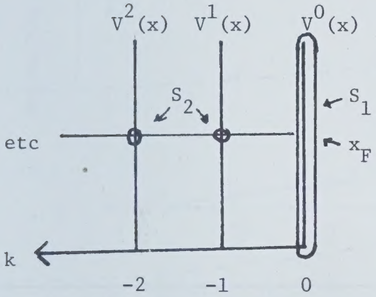
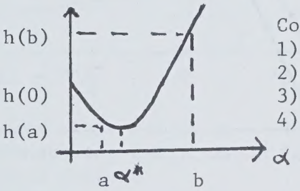
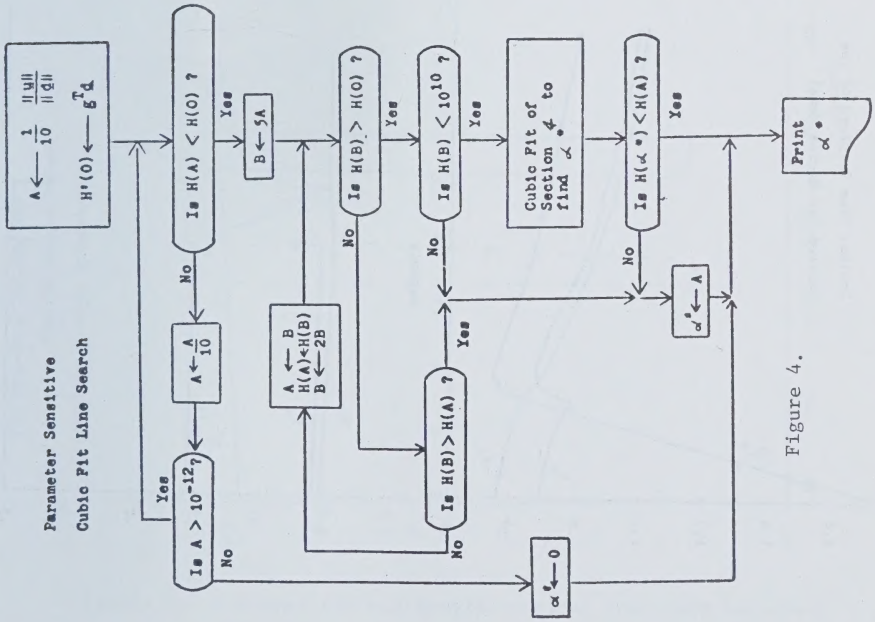

$V^2(x)$
$V^1(x)$
$V^0(x)$
S_2
S_1
x_F
etc
k
-2
-1
0

Figure 1.



Conditions
1) $h'(0) < 0$
2) $h(a) < h(0)$
3) $h(b) > h(0)$
4) $0 < a < b$

Figure 3.


$A \leftarrow \frac{1}{10} \frac{\|y\|}{\|d\|}$
$H'(0) \leftarrow g^T d$
Is $H(A) < H(0)$?
Yes
No
$B \leftarrow 5A$
Is $H(B) > H(0)$?
Yes
No
Is $H(B) < 10^{10}$?
Yes
No
Cubic Fit of Section 4 to find α^*
Is $H(\alpha^*) < H(A)$?
Yes
No
Print α^*
$A \leftarrow \frac{A}{10}$
$A \leftarrow B$
$H(A) \leftarrow H(B)$
$B \leftarrow 2B$
Is $H(B) > H(A)$?
Yes
No
$\alpha^* \leftarrow A$
Is $A > 10^{-12}$?
Yes
No
$\alpha^* \leftarrow 0$

Parameter Sensitive
Cubic Fit Line Search

Figure 4.

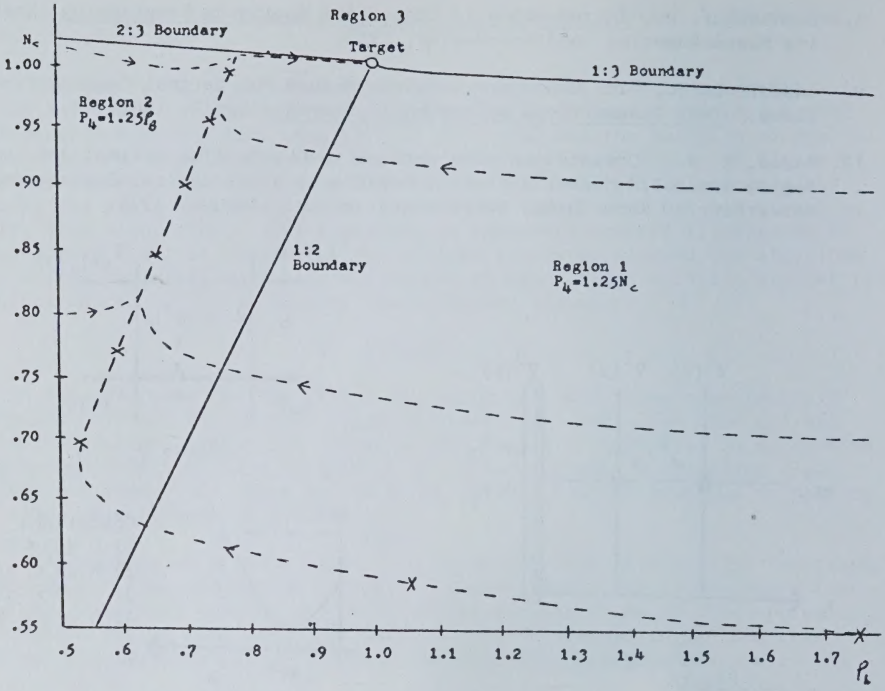


Figure 5. State set motion when optimal control is applied.

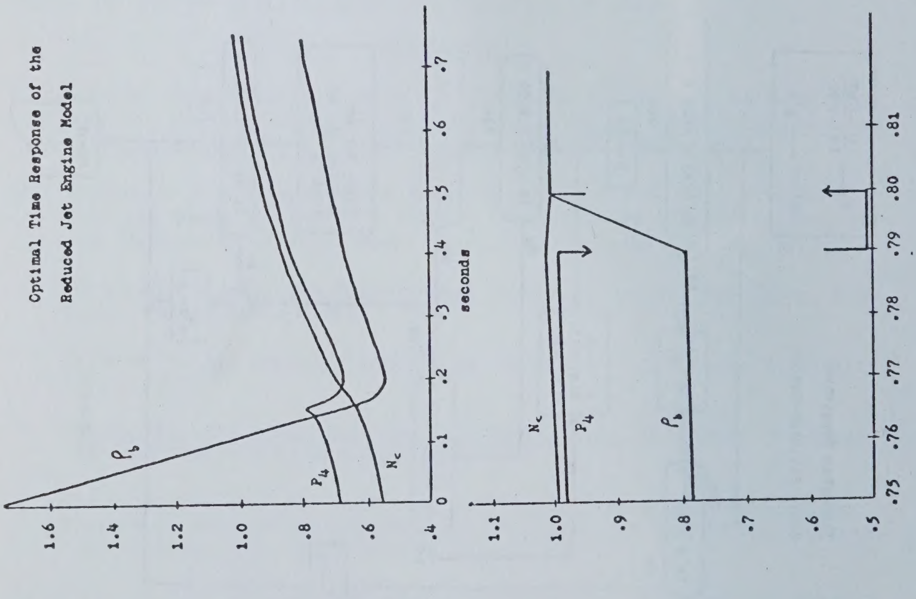


Figure 6.

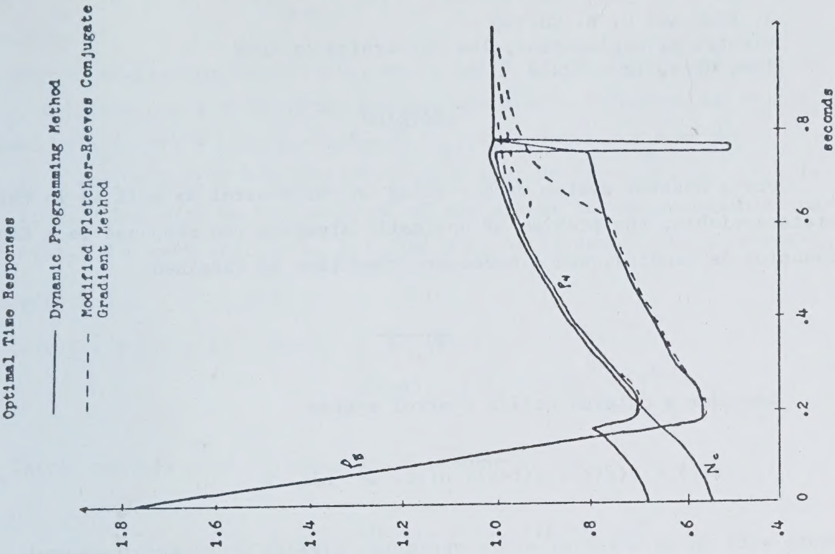


Figure 7. A comparison of the optimal trajectories produced by the two computational methods.

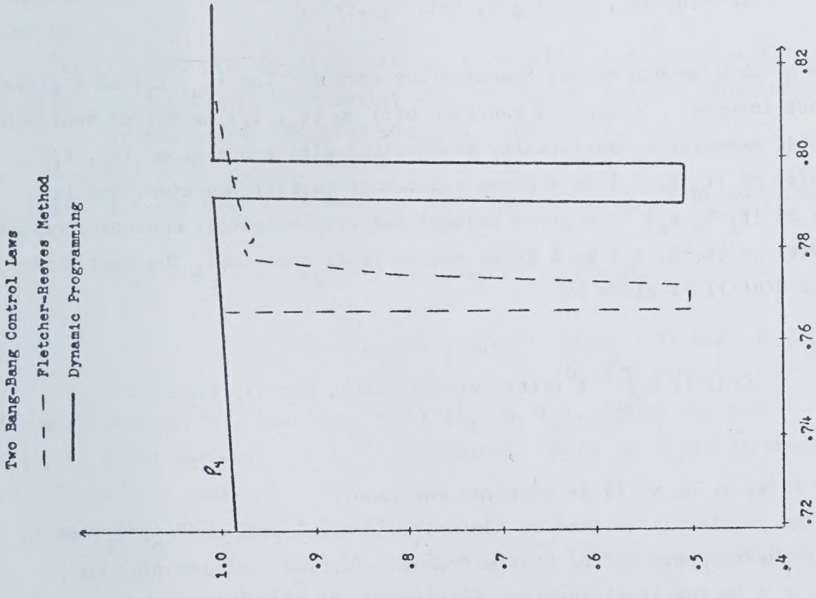


Figure 8. A comparison of the time optimal open loop bang-bang controls.

OPTIMAL CONTROL OF DELAYED ACTION CONTROL SYSTEM IN FUNCTION SPACE

Z. BIEN and D. H. CHYUNG
Electrical Engineering, The University of Iowa
Iowa City, Iowa 52242

ABSTRACT

For a control system with a delay in the control as well as in the state variable, the problem of optimally steering the response to a final function is studied, and a necessary condition is obtained.

Consider a delayed action control system

$$\dot{x}(t) = f(x(t), x(t-h), u(t), u(t-h), t)$$

where x(t) is an n-vector state variable, u(t) is an m-vector control variable, and $h > 0$ is the time delay. Let Ω be the control restraint set given by

$$\Omega = \{u | u \varepsilon R^m, q^i(u) \leqq 0, i=1, \ldots, r\} ,$$

where q^i is a smooth scalar function for each i. Let $[t_o, t_1]$ be a given compact interval. A control function u(t) on $[t_o, t_1]$ is called admissible if it is measurable, essentially bounded and $u(t) \varepsilon \Omega$ a.e. on $[t_o, t_1]$. Let $\phi(t)$ on $[t_o-h, t_o]$ be a given continuous initial function, and let $u_o(t)$ on $[t_1-h, t_o]$ be a given bounded measurable initial control function. Let $\psi(t)$ on $[t_1-h, t_1]$ be a given smooth final function. The cost functional $C(u(\cdot))$ is given by

$$C(u(\cdot)) = \int_{t_o}^{t_1} f^o(x(t), x(t-h), u(t), u(t-h), t)dt$$

where $f^o(x, y, u, v, t)$ is a scalar function.

The problem is to find an admissible control u(t) on $[t_o, t_1]$ which steers the response of the system from the initial function $\phi(t)$ on $[t_o-h, t_o]$ to the final function $\psi(t)$ on $[t_1-h, t_1]$ while minimizing the cost functional $C(u(\cdot))$.

The partial derivatives are denoted by subscripts, e.g., $f_x(x_o, y, u, v, t) = \frac{\partial}{\partial x} f(x, y, u, v, t)\big|_{x=x_o}$, etc. Given a matrix B, the transpose is denoted by B^T.

For an admissible control u(t) on I, let $A_o(u(\cdot))$ denote the cone in $L^m_\infty[t_o, t_1]$ (the space of m-vector bounded measurable functions on $[t_o, t_1]$) defined by $A_o(u(\cdot)) = \{\alpha\delta u | \alpha > 0, \delta u \varepsilon L^m_\infty [t_o, t_1], \delta u(t) = 0$ a.e. on $[t_o, t_1 - 2h]$, $q^i(u(t)) + q^i_u(u(t))\delta u(t) < 0$ a.e. on $[t_1 - 2h, t_1]$, $i = 1, 2, \ldots, r\}$. Let $\bar{x}(t)$ be the response of the system for an admissible control $\bar{u}(t)$. For each $t \varepsilon [t_1 - h, t_1]$, let N(t, u(t)) denote the n × n matrix

$$\begin{aligned} &N(t, u(t)) \\ &= f_u(\bar{x}(t), \bar{x}(t-h), u(t), u(t-h), t)\, [f_u(\bar{x}(t), \\ &\qquad\qquad \bar{x}(t-h), u(t), u(t-h), t)]^T \\ &+ f_v(\bar{x}(t), \bar{x}(t-h), u(t), u(t-h), t)\, [f_v(\bar{x}(t), \\ &\qquad\qquad \bar{x}(t-h), u(t), u(t-h), t)]^T. \end{aligned}$$

Define the n × n matrix function $\tilde{N}(t, u(t))$ on $[t_1 - h, t_1]$ by $\tilde{N}(t, u(t)) = N(t, u(t))^{-1}$ if N(t, u(t)) is nonsingular at t and $\tilde{N}(t, u(t)) = 0$ otherwise. An admissible control u(t) is called regular with respect to the response $\bar{x}(t)$ if

(i) $f(\bar{x}(t), \bar{x}(t-h), u(t), u(t-h), t) = \dot{\psi}(t)$ a.e. on $[t_1 - h, t_1]$,

(ii) N(t, u(t)) is nonsingular a.e. on $[t_1 - h, t_1]$ and each component of the matrix function $\tilde{N}(t, u(t))$ on $[t_1 - h, t_1]$ is in $L_\infty[t_1 - h, t_1]$,

(iii) for any $z \varepsilon L^n_\infty [t_1 - h, t_1]$, there exists $\delta u_z \varepsilon A_o(u(\cdot))$ such that

$$\begin{aligned} z(t) &= f_u(\bar{x}(t), \bar{x}(t-h), u(t), u(t-h), t)\delta u_z(t) \\ &\quad + f_v(\bar{x}(t), \bar{x}(t-h), u(t), u(t-h), t)\delta u_z(t-h) \text{ a.e. on } [t_1 - h, t_1] \end{aligned}$$

For each $(x,y) \varepsilon R^n \times R^n$, $\hat{u} \varepsilon \Omega$, $\hat{v} \varepsilon \Omega$, and $t \varepsilon [t_1 - h, t_1]$, define the sets $W_1(x, y, \hat{v}, t)$ and $W_2(x, y, \hat{u}, t)$ as follows. $W_1(x, y, \hat{v}, t)$ is the set of all $u \varepsilon \Omega$ such that

(i) $f(x, y, u, \hat{v}, t) = \dot{\psi}(t)$,

(ii) $f_u(x, y, u, \hat{v}, t) f_u(x, y, u, \hat{v}, t)^T + f_v(x, y, u, \hat{v}, t) f_v(x, y, u, \hat{v}, t)^T$ is nonsingular,

(iii) $\{f_u(x, y, u, \hat{v}, t)\Delta u + f_v(x, y, u, \hat{v}, t)\Delta v | \Delta u \varepsilon R^m, q^j_u(j)\Delta u < 0$

for each j=1, ..., r such that $q^j(u) = 0$, $\Delta v \varepsilon R^m$, $q^j_u(\hat{v})\Delta v < 0$ for each j=1, ..., r such that $q^j(\hat{v}) = 0\} = R^n$

Let $W_2(x, y, \hat{u}, t)$ be the set of all $v \varepsilon \Omega$ such that

(i) $f(x, y, \hat{u}, v, t) = \dot{\psi}(t)$

(ii) $f_u(x, y, \hat{u}, v, t) f_u(x, y, \hat{u}, v, t)^T + f_v(x, y, \hat{u}, v, t) f_v(x, y, \hat{u}, v, t)^T$ is nonsingular,

(iii) $\{f_u(x, y, \hat{u}, v, t)\Delta u + f_v(x, y, \hat{u}, v, t)\Delta v \mid \Delta u \varepsilon R^m$, $q^j_u(\hat{u})\Delta u < 0$ for each j=1, ..., r such that $q^j(u) = 0$, $\Delta v \varepsilon R^m$, $q^j_u(v)\Delta v < 0$ for each j=1, ..., r such that $q^j(v) = 0\} = R^n$.

For notational convenience, let $V_1^*(t) = W_1(x^*(t), x^*(t-h), u^*(t-h), t)$, $V_2^*(t) = W_2(x^*(t+h), x^*(t), u^*(t+h), t+h)$, $\hat{\eta}(t) = (\eta^o(t), \eta(t))$ and $\hat{f}(x, y, u, v, t) = (f^o(x, y, u, v, t), f(x, y, u, v, t))$.

By extending the result in [3], the following necessary condition for an optimal control can be obtained.

NECESSARY CONDITION. Let $u^*(t)$ with response $x^*(t)$ be an optimal control and let $u^*(t)$ be regular with respect to $x^*(t)$. Then there exist vector functions $\mu(t) = (\mu^1(t), \ldots, \mu^r(t))$ with each component in $L_\infty[t_1-2h, t_1]$ and $\nu(t) = (\nu^1(t), \ldots, \nu^n(t))$ with each component in $L_\infty[t_1-h, t_1]$, and absolutely continuous functions $\eta^o(t)$, $\eta(t) = (\eta^1(t), \eta^2(t), \ldots, \eta^n(t))$ on $[t_o, t_1]$ such that

$$(\eta^o(t_1), \eta(t_1)) \neq 0$$

$$\dot{\eta}^o(t) = 0, \quad t_o \leqq t \leqq t_1, \quad \eta^o = \eta^o(t_1) \leqq 0$$

$$\begin{aligned}
-\dot{\eta}(t) = \eta^o[&f^o_x(x^*(t), x^*(t-h), u^*(t), u^*(t-h), t) \\
&+ f^o_y(x^*(t+h), x^*(t), u^*(t+h), u^*(t), t+h)] \\
&+ \eta(t) f_x(x^*(t), x^*(t-h), u^*(t), u^*(t-h), t) \\
&+ \eta(t+h) f_y(x^*(t+h), x^*(t), u^*(t+h), u^*(t), t+h)
\end{aligned}$$

$$\text{a.e. on } [t_o, t_1-2h],$$

$$\begin{aligned}
-\dot{\eta}(t) = \eta^o[&f^o_x(x^*(t), x^*(t-h), u^*(t), u^*(t-h), t) \\
&+ f^o_y(x^*(t+h), x^*(t), u^*(t+h), u^*(t), t+h)]
\end{aligned}$$

$$+ \quad \eta(t)f_x(x^*(t),\ x^*(t-h),\ u^*(t),\ u^*(t-h),\ t)$$

$$+ \quad [\eta(t+h) + \nu(t+h)]f_y(x^*(t+h),\ x^*(t),\ u^*(t+h),\ u^*(t),\ t)$$

$$\text{a.e. on } [t_1-2h,\ t_1-h],$$

$$-\dot{\eta}(t) = \eta^o f^o_x(x^*(t),\ x^*(t-h),\ u^*(t),\ u^*(t-h),\ t)$$

$$+ [\eta(t) + \nu(t)]f_x(x^*(t),\ x^*(t-h),\ u^*(t),\ u^*(t-h),\ t)$$

$$\text{a.e. on } [t_1-h,\ t_1].$$

$$\hat{\eta}(t)\hat{f}(x^*(t),\ x^*(t-h),\ u^*(t),\ u^*(t-h),\ t)$$

$$+ \hat{\eta}(t+h)\hat{f}(x^*(t+h),\ x^*(t),\ u^*(t+h),\ u^*(t),\ t+h)$$

$$= \max_{u\,\varepsilon\,\Omega} \ [\hat{\eta}(t)\hat{f}(x^*(t),\ x^*(t-h),\ u,\ u^*(t-h),\ t)$$

$$+ \hat{\eta}(t+h)\hat{f}(x^*(t+h),\ x^*(t),\ u^*(t+h),\ u,\ t+h)]$$

$$\text{a.e. on } [t_o,\ t_1-2h],$$

$$\hat{\eta}(t)\hat{f}(x^*(t),\ x^*(t-h),\ u^*(t),\ u^*(t-h),\ t)$$

$$+ \hat{\eta}(t+h)\hat{f}(x^*(t+h),\ x^*(t),\ u^*(t+h),\ u^*(t),\ t+h)$$

$$= \max_{v \varepsilon V^*_2(t)} \ [\hat{\eta}(t)\hat{f}(x(t),\ x(t-h),\ v,\ u^*(t-h),\ t)$$

$$+ \hat{\eta}(t+h)\hat{f}(x^*(t+h),\ x^*(t),\ u^*(t+h),\ v,\ t+h)]$$

$$\text{a.e. on } [t_1-2h,\ t_1-h],$$

$$\hat{\eta}(t)\hat{f}(x^*(t),\ x^*(t-h),\ u^*(t),\ u^*(t-h),\ t)$$

$$= \max_{u \varepsilon V^*_1(t)} \ [\hat{\eta}(t)\hat{f}(x^*(t),\ x^*(t-h),\ u,\ u^*(t-h),\ t)]$$

$$\text{a.e. on } [t_1-h,\ t_1],$$

$$\mu(t)q_u(u^*(t)) + \nu(t+h)f_v(x^*(t+h),\ x^*(t),\ u^*(t+h),\ u^*(t),\ t+h)$$

$$+ \hat{\eta}(t)\hat{f}_u(x^*(t),\ x^*(t-h),\ u^*(t),\ u^*(t-h),\ t)$$

$$+ \hat{\eta}(t+h)\hat{f}_v(x^*(t+h),\ x^*(t),\ u^*(t+h),\ u^*(t),\ t+h) = 0$$

$$\text{a.e. on } [t_1-2h,\ t_1-h],$$

$$\mu(t)q_u(u^*(t)) + \nu(t)f_u(x^*(t),\ x^*(t-h),\ u^*(t),\ u^*(t-h),\ t)$$

$$+ \hat{\eta}(t)\hat{f}_u(x^*(t),\ x^*(t-h),\ u^*(t),\ u^*(t-h),\ t) = 0$$

$$\text{a.e. on } [t_1-h,\ t_1],$$

$$\mu^i(t) \leqq 0\ ,\quad \text{a.e. on } [t_1-2h,\ t_1],\ i=1,\ \ldots,\ r,$$

$$\mu^i(t)q^i(u^*(t))) = 0\ ,\quad \text{a.e. on } [t_1-2h,\ t_1],\ i=1,\ \ldots,\ r.$$

The above result is also sufficient for certain class of linear systems. The case when some components of the final function are free can be handled in a similar manner.

REFERENCES

1. Banks, H. T., Necessary Conditions for Control Problems with Variable Time Lags, SIAM J. Control, Vol. 6, No. 7, 1968.
2. Banks, H. T., and A. Manitus, Application of Abstract Variational Theory to Hereditary Systems - A Survey, IEEE Transactions on Automatic Control, Vol. AC-19, pp 524-533.
3. Bien, Z. and D. H. Chyung, Optimal Control of Delay Systems, Proceedings of 13th Annual Allerton Conference on Circuit and Systems Theory, University of Illinois, October, 1975.
4. Chyung, D. H. and E. B. Lee, Delayed Action Control Problem, Automatica Vol. 6, No. 3, May, 1970.
5. Jacobs, M. Q. and T-J. Kao, An Optimum Settling Problem for Time Lag System, J. of Math. Anal. & Appli., Vol. 40, No. 3, Dec., 1972.
6. Makowski, K. and L. W. Neustadt, Optimal Control Problems with Mixed Control-Phase Variable Equality and Inequality Constraints, SIAM J. Control Vol. 12, No. 2, May, 1974.

A NEW ANALYTICAL REPRESENTATION FOR SECTION-WISE PIECEWISE-LINEAR FUNCTIONS AND ITS APPLICATIONS

LEON O. CHUA
Dept. of Electrical Engineering
and Computer Sciences
and the Electronics Research Laboratory
University of California, Berkeley
Berkeley, California 94720

SUNG MO KANG
Dept. of Electrical Engineering
Rutgers University
New Brunswick, New Jersey 08903

ABSTRACT

This paper presents a new closed form analytical formula for representing n-dimensional surfaces and scalar functions of n variables which are piecewise-linear over each cross section obtained by freezing any combination of n-1 of the n coordinates. This new section-wise piecewise-linear representation can be easily programmed with efficient computer storage. As a global representation, it can be used to derive explicit closed form expressions for system parameters and design formulas. Examples are given which illustrate the potential applications of this representation in the modeling and analysis of nonlinear devices, circuits and systems.

INTRODUCTION

Piecewise-linear approximation of one or more cross sections of nonlinear multi-dimensional characteristics and surfaces has been widely used in many fields of science and engineering [1-7].

The current piecewise-linear approximation technique, however, suffers from three major drawbacks: (1) the lack of an explicit analytical representation, (2) the need to store an immense amount of data for high-dimensional functions for computational purposes, and (3) high overhead cost involved in programming and bookkeeping operations involving piecewise-linear functions.

Our objective in this paper is to present an explicit analytical representation for multi-dimensional function $f(x_1,x_2,\ldots,x_n)$ which is piecewise-linear over each cross-section in $\mathbb{R}^n$ defined by holding any n-1 coordinates fixed. We call this a section-wise piecewise-linear representation to emphasize the fact that while our representation agrees with the conventional piecewise-linear representation for functions of one variable (n=1), it is not a piecewise-linear representation when n>1. On the contrary, our canonical representation for $n \geq 2$ is at least quadratic in the sense that it contains all product term combinations such as $x_j, x_jx_k, x_jx_kx_\ell, \ldots, x_jx_k\ldots x_n$. Since our representation is expressed in analytic form, it is a global representation requiring simple programming efforts while allowing efficient data storage and retrieval. But above all, our closed form representation allows standard mathematical operations and manipulations to be performed on equations involving section-wise piecewise-linear representations. In particular, explicit coefficients, parameters, and design formulas for many practical problems can now be derived in closed form.

The following material is divided into two parts. Part I presents the closed form canonical formulas for representing any section-wise piecewise-linear functions while Part II presents some properties and applications of these representations.

Research sponsored in part by the Office of Naval Research Contract N0014-76-C-0572 and the National Science Foundation Grant ENG75-22282.

I. CANONICAL REPRESENTATIONS

A. A Canonical Piecewise-Linear Representation for Single-Valued Functions with Finite Jump Discontinuities

A typical piecewise-linear function with finite jump discontinuities is shown in Fig. 1. We will always label the segments consecutively from "0" (leftmost segment) through "n" (rightmost setment) and let "m_j" denote the slope of segment j. Corresponding to the n+1 segments, partition the x-axis into n+1 intervals $I_0 \triangleq (-\infty,x_1]$, $I_1 \triangleq (x_1,x_2],\ldots,I_{n-1} \triangleq (x_{n-1},x_n]$, and $I_n \triangleq (x_n,\infty)$, as shown in Fig. 1(b).

Main Theorem: Canonical Piecewise-Linear Representation

Any single-valued piecewise-linear function with at most n finite jump discontinuities at the n breakpoints $x_1 < x_2 < \ldots < x_n$ can be represented uniquely by[1]

$$f(x) = a_0 + a_1 x + \sum_{j=1}^{n} \left\{ b_j |x-x_j| + c_j \,\mathrm{sgn}(x-x_j) \right\} \tag{1}$$

where the coefficients are given explicitly by:[2]

$$a_1 = \frac{1}{2}(m_0+m_n) \tag{2}$$

$$b_j = \frac{1}{2}(m_j-m_{j-1}),\ j = 1,2,\ldots,n \tag{3}$$

$$c_j = \begin{cases} 0, & \text{if } f(\cdot) \text{ is continuous at the breakpoint } x = x_j \\ \frac{1}{2}\left[f(x_j^+) - f(x_j^-)\right], & \text{otherwise} \end{cases} \tag{4}$$

$$a_0 = f(0) - \sum_{j=1}^{n} \left(b_j |x_j| - c_j \,\mathrm{sgn}(x_j) \right) \tag{5}$$

The proof of this theorem is given in reference 8.

B. A Canonical Piecewise-Linear Representation for a Family of Single-Valued Functions with Jump Discontinuities

The dc characteristic curves of 3-terminal devices, such as transistors, FET's, etc., are generally modeled by a set of input characteristics and a set of output characteristics, respectively [5,9]. Each set is usually measured by a curve tracer [10] and, for large-signal applications, the characteristics can often be approximated realistically by a family of piecewise-linear curves

$$y = f(x,p) \tag{6}$$

where p denotes a parametric variable.

Let $\mathscr{F} = \{y = f(x,p)\colon p = p_1,p_2,\ldots,p_N\}$ denote a family of single-valued piecewise-linear curves having finite jump discontinuities. Without loss of generality, we can assume each piecewise-linear curve $f(x,p_i)$ has n breakpoints $x_1(p_i) < x_2(p_i) < \ldots < x_n(p_i)$, and n+1 segments with slopes $m_0(p_i)$, $m_1(p_i),\ldots,m_n(p_i)$, respectively. Such a curve can obviously be represented by (1). Now if we let p assume the N values assigned to the given curves in $\mathscr{F}$, we can represent each curve in $\mathscr{F}$ exactly by the following section-wise piecewise-linear canonical representation:

---

[1] We define $|x| \triangleq \begin{cases} x, & x > 0 \\ -x, & x \leq 0 \end{cases}$ and $\mathrm{sgn}(x) \triangleq \begin{cases} 1, & x > 0 \\ -1, & x \leq 0 \end{cases}$

[2] Recall m_j denotes the slope of segment j.

$$f(x,p) = a_0(p) + a_1(p)x + \sum_{j=1}^{n} \left\{ b_j(p)\cdot |x-x_j(p)| + c_j(p)\,\mathrm{sgn}\left(x-x_j(p)\right)\right\} \tag{7}$$

where $a_0(\cdot)$, $a_1(\cdot)$, $b_j(\cdot)$, and $c_j(\cdot)$ are similarly defined as in (2)-(5). Observe that (7) is completely specified by at most 3n+2 functions of a single variable p; namely, $a_0(p)$, $a_1(p)$, $b_j(p)$, $c_j(p)$, and $x_j(p)$, $j = 1,2,\ldots,n$, where n is the number of breakpoints in each curve in $\mathcal{F}$. Moreover, if all curves in $\mathcal{F}$ are continuous, then $c_j(p) \equiv 0$ and only 2n+2 functions are needed. These functions may be represented by any convenient interpolation scheme, such as Lagrange or Hermite polynomial [11], or by any convenient analytical formula such as the piecewise-linear representation given in (1).

Example 1. Consider the anode-to-cathode dc characteristic curves of a typical silicon-controlled rectifier (SCR) as shown in Fig. 2(a) and its 5-segment piecewise-linear approximation as shown in Fig. 2(b). In this case, $\mathcal{F} = \{V_A = f(I_A,I_g)\colon\ I_g = 0,5,10,15,20\}$, where the gate current I_g is the parametric variable. Observe that n=4 since each curve has 4 breakpoints, and N=5 since $\mathcal{F}$ has 5 curves corresponding to 5 values of I_g. Since all curves in Fig. 2(b) are continuous, $c_j(I_g) \equiv 0$ and we can model the given family of curves exactly by

$$V_A = f(I_A,I_g) = a_0(I_g) + a_1(I_g)I_A + b_1(I_g)\cdot|I_A-I_1(I_g)| + b_2(I_g)\cdot |I_A-I_2(I_g)| + b_3(I_g)\cdot|I_A-I_3(I_g)| + b_4(I_g)\cdot|I_A-I_4(I_g)| \tag{8}$$

where $I_1(I_g)$, $I_2(I_g)$, $I_3(I_g)$, and $I_4(I_g)$ denote the 4 breakpoint locations for each value of $I_g = 0,5,10,15,20$, as shown in Figs. 2(c)-(f). Observe that for simplicity, the 4 breakpoints in Figs. 2(c)-(f) are connected by straight line segments, thereby allowing each breakpoint function $I_j(I_g)$ to be represented by (1).

Since n=4 and since $c_j(I_g) \equiv 0$, (8) requires only 2n+2 = 10 model functions; namely, $a_0(I_g)$, $a_1(I_g)$, $b_j(I_g)$ and $I_j(I_g)$, $j = 1,2,3,4$. The functions $a_0(\cdot)$, $a_1(\cdot)$ and $b_j(\cdot)$ are shown in Figs. 2(g)-(1). Again, for complete generality, we represent these functions by piecewise-linear curves so that they can in turn be represented by (1).

The preceding representations are valid only for single-valued functions. A generalization of these representations to an important class of multi-valued piecewise-linear curves is given in reference 8.

C. A Canonical Sectionwise-Piecewise-Linear Representation for Continuous Functions of Several Variables

Our objective in this section is to generalize the basic approach presented in Section I-B and develop a canonical representation for continuous functions $y = f(x_1,x_2,\ldots,x_n)$ of n variables. We assume a priori that a sufficient number of data points has been taken such that the function can be realistically approximated by a piecewise-linear function of a single variable along every cross-section obtained by freezing any combination of (n-1) of the n coordinates.

Consider first the case n=3 and suppose that the values of $f(x_1, x_2, x_3)$ are given at a set of N^3 data points (x_{1i},x_{2j},x_{3k}), $i,j,k = 1,2,\ldots,N$. By assumption, the function $f(x_1,x_{2j},x_{3k})$ with the second and third coordinates fixed at $x_2 = x_{2j}$ and $x_3 = x_{3k}$ is a piecewise-linear continuous function of a single variable x_1 and can therefore be represented by (1); namely,[3]

---

[3] The upper index of summation in (9) is N-2, rather than N because the left-most and the rightmost data points are needed to compute for the slope of the two end segments, and hence only N-2 data points are available as breakpoints.

$$f(x_1,x_{2j},x_{3k}) = a_0(x_{2j},x_{3k}) + a_1(x_{2j},x_{3k})x_1 + \sum_{i=1}^{N-2} b_i(x_{2j},x_{3k})|x_1-x_{1i}| \quad (9)$$

where the coefficients a_0, a_1 and b_i are determined via (5), (2) and (3), respectively. Notice that these coefficients will change for a different choice of (x_{2j},x_{3k}) and hence can be considered as functions of the two variables x_2 and x_3 whose values at any data point (x_{2j},x_{3k}) can be computed from (5), (2) and (3). Since $a_0(\cdot)$, $a_1(\cdot)$, and $b_i(\cdot)$ are now functions of two instead of the original three variables, they can be represented in turn by the canonical representation (7). It follows from the preceding algorithm that any continuous function $f(x_1,x_2,x_3)$ of three variables having a piecewise-linear cross section can be represented by the following canonical form

$$f(x_1,x_2,x_3) = a_0(x_2,x_3) + a_1(x_2,x_3)x_1 + \sum_{i=1}^{N-2} b_i(x_2,x_3)|x_1-x_{1i}| \quad (10)$$

where $a_0(x_2,x_3) = a_0^{a_0}(x_3) + a_1^{a_0}(x_3)x_2 + \sum_{i=1}^{N-2} b_i^{a_0}(x_3)|x_2-x_{2i}| \quad (11)$

$$a_1(x_2,x_3) = a_0^{a_1}(x_3) + a_1^{a_1}(x_3)x_2 + \sum_{i=1}^{N-2} b_i^{a_1}(x_3)|x_2-x_{2i}| \quad (12)$$

$$b_j(x_2,x_3) = a_0^{b_j}(x_3) + a_1^{b_j}(x_3)x_2 + \sum_{i=1}^{N-2} b_i^{b_j}(x_3)|x_2-x_{2i}| \quad (13)$$

where the double superscripts attached to the coefficients $a_0(x_3)$, $a_1(x_3)$, and $b_j(x_3)$ are used to identify the associated model function. Observe that the model functions in (11)-(13) are now all functions of a single variable x_3, and hence can in turn be represented by (1).

The generalization of the preceding algorithm to continuous functions of any number "n" of variables is now obvious. We first freeze all (n-1) coordinates $x_2,x_3,\ldots,x_n$ and write

$$f(x_1,x_2,x_3,\ldots,x_n) = a_0(x_2,x_3,\ldots,x_n) + a_1(x_2,x_3,\ldots,x_n)x_1 + \sum_{j=1}^{N-2} b_j(x_2,x_3,\ldots,x_n)|x_1-x_{1j}| \quad (14)$$

where $a_0(x_2,x_3,\ldots,x_n)$, $a_1(x_2,x_3,\ldots,x_n)$, and $b_j(x_2,x_3,\ldots x_n)$ are functions of n-1 variables, which is one less than the original number of variables. The same algorithm can be applied repeatedly to these model functions, where the number of variables is reduced by one after each iteration. The algorithm must clearly terminate when all model functions have been reduced to functions of a single variable, which in turn are represented by (1).

II. PROPERTIES AND APPLICATIONS

A. Properties and Manipulations of Piecewise-Linear Functions

The canonical piecewise-linear representation (1) has been shown in Part I to be the fundamental building block from which all section-wise piecewise-linear representations for functions of several variables are based. It is desirable therefore that we uncover as many properties as possible with respect to the constraints on the coefficients a_0,a_1,b_j, and c_j. To simplify our notations and derivations, unless otherwise stated, we will assume the function is continuous so that $c_j=0$ and (1) reduces to the form:

$$f(x) = a_0 + a_1x + \sum_{j=1}^{n} b_j|x-x_j| \quad (15)$$

The proofs of the following properties are given in reference 8.

Property 1. If the breakpoints of $f(x)$ are such that $0 < x_1 < x_2 \ldots < x_n$, then

$$a_0 = \frac{1}{2}[f(0) + f(x_n) - m_n x_n] \tag{16}$$

Property 2. A continuous piecewise-linear function $f(x)$ is *monotone-increasing* (resp., strictly monotone-increasing) if, and only if,

$$a_1 + \sum_{j=1}^{k} b_j - \sum_{j=k+1}^{n} b_j \geq 0 \text{ (resp., } > 0) \text{ for all } k = 0,1,2,\ldots,n \tag{17}$$

Property 3. If a continuous piecewise-linear function has *monotone-increasing slopes* (resp., *monotone-decreasing slopes*) in the sense that $0 \leq m_0 < m_1 < \ldots < m_n$ (resp., $m_0 > m_1 \ldots > m_n \geq 0$), then

(1) $a_1 > 0$

(2) $b_j > 0$ (resp., $b_j < 0$) for all $j = 1,2,\ldots,n$.

(3) if $f(0) = 0$, then $a_0 \leq 0$ (resp., $a_0 \geq 0$)

Property 4. Let $y = f(x) = a_0 + a_1 x + \sum_{j=1}^{n} b_j |x-x_j|$ be a continuous and *strictly-increasing* piecewise-linear function and let its *inverse* be given by

$$x = f^{-1}(y) = \hat{a}_0 + \hat{a}_1 y + \sum_{j=1}^{n} \hat{b}_j |y-y_j| \tag{18}$$

then the coefficients in (18) can be computed as follows:

$$\hat{a}_1 = a_1 \Big/ \left[a_1^2 - \left(\sum_{j=1}^{n} b_j \right)^2 \right] \tag{19}$$

$$\hat{b}_k = -b_k \Big/ \left[\left(a_1 + \sum_{j=1}^{k-1} b_j - \sum_{j=k+1}^{n} b_j \right)^2 - b_k^2 \right], \quad k = 1,2,\ldots,n \tag{20}$$

$$\hat{a}_0 = f^{-1}(0) - \sum_{j=1}^{n} \hat{b}_j |f(x_j)| \tag{21}$$

Remark. Property 4 can be generalized to allow finite jump discontinuities in $f(x)$. To be more specific, suppose $f(x)$ is continuous except at breakpoint x_k, where $y_k^- = f(x_k^-)$ and $y_k^+ = f(x_k^+)$. Then

$$y = f(x) = a_0 + a_1 x + \sum_{j=1}^{n} b_j |x-x_j| + c_k \operatorname{sgn}(x-x_k) \tag{22}$$

and the *inverse function* is given by

$$x = f^{-1}(y) = \tilde{a}_0 + \tilde{a}_1 y + \sum_{\substack{j=1 \\ j \neq k}}^{n} \tilde{b}_j |y-y_j| + \tilde{b}_k^- |y-y_k^-| + \tilde{b}_k^+ |y-y_k^+| \tag{23}$$

It can be shown that the coefficients $\tilde{a}_1$ and $\tilde{b}_j$, $j \neq k$ in (23) can be computed from the same formulas given in (19) and (20), and that

$$\tilde{b}_k^- = -\frac{1}{2} \Big/ \left[a_1 + \sum_{j=1}^{k-1} b_j - \sum_{j=k}^{n} b_j \right] \tag{24}$$

$$\tilde{b}_k^+ = \frac{1}{2} \Big/ \left[a_1 + \sum_{j=1}^{k} b_j - \sum_{j=k+1}^{n} b_j \right] \tag{25}$$

$$\tilde{a}_0 = f^{-1}(0) - \left\{ \sum_{\substack{j=1 \\ j \neq k}}^{n} \tilde{b}_j |y_j| + \tilde{b}_k^- |y_k^-| + \tilde{b}_k^+ |y_k^+| \right\} \tag{26}$$

Property 5. Let $y = f(x)$ be any piecewise-linear curve and let $x = g(z)$ be any strictly monotone increasing piecewise-linear curve, then their composition $y \triangleq f \circ g(z) = f(g(z)) = h(z)$ is also a piecewise-linear curve and the coefficients characterizing $h(z)$ can be computed explicitly from those characterizing $f(x)$ and $g(z)$. In particular, if

$$y = f(x) = a_0^f + a_1^f x + \sum_{j=1}^{n_f} \left\{ b_j^f |x - x_j| + c_j^f \operatorname{sgn}(x - x_j) \right\} \tag{27}$$

$$x = g(z) = a_0^g + a_1^g z + \sum_{j=1}^{n_g} \left\{ b_j^g |z - z_j| + c_j^g \operatorname{sgn}(z - z_j) \right\} \tag{28}$$

then the composition function $f \circ g(\cdot)$ is given explicitly by:

$$\begin{aligned} y = h(z) = {} & a_0^f + a_1^f a_0^g + \sum_{i=1}^{n_f} b_i^f \left[\sum_{j=1}^{n_g} \left(c_j^g - b_j^g z_j \right) \right] + \left[a_1^f a_1^g + \sum_{i=1}^{n_f} b_i^f \left(\sum_{j=1}^{n_g} b_j^g \right) \right] z \\ & + \sum_{j=1}^{n_g} b_j^g \left(a_1^f + \sum_{i=1}^{n_f} b_i^f \operatorname{sgn}(z_j - z_i^*) \right) |z - z_j| \\ & + \sum_{i=1}^{n_f} b_i^f \left(a_1^g - \sum_{j=1}^{n_g} b_j^g \operatorname{sgn}(z_j - z_i^*) \right) |z - z_i^*| \\ & + \sum_{j=1}^{n_g} c_j^g \left(a_1^f + \sum_{i=1}^{n_f} b_i^f \operatorname{sgn}(z_j - z_i^*) \right) \operatorname{sgn}(z - z_j) \\ & + \sum_{i=1}^{n_f} \left\{ b_i^f \left[a_0^g - x_i + a_1^g z_i^* + \sum_{j=1}^{n_g} \left(b_j^g |z_j - z_i^*| - c_j^g \right. \right. \right. \\ & \left. \left. \left. \operatorname{sgn}(z_j - z_i^*) \right) \right] + c_i^f \right\} \operatorname{sgn}(z - z_i^*) \end{aligned} \tag{29}$$

where z_i^* is the unique solution of $g(z) - x_i = 0$, $i = 1, 2, \ldots, n_f$.

B. Some Applications of Canonical Piecewise-Linear Representations

The preceding properties are particularly useful in deriving algebraically both the DP (driving-point) and TC (transfer characteristics) plots of resistive nonlinear networks. Indeed, much of the graphical procedures used for deriving DP and TC plots in [5] can now be replaced by algebraic manipulations, as the next example demonstrates.

Example 2. Consider the nonlinear voltage divider circuit shown in Fig. 3(a) where the nonlinear resistors R_1 and R_2 are characterized by the piecewise-linear curves shown in Figs. 3(c) and (d), respectively. The problem is to derive the v_0-vs.-v_i TC plot.

Applying KVL and KCL, we obtain

$$v_i = f \circ g(v_0) + v_0 \triangleq F(v_0) \tag{30}$$

Now applying (1) to the v-i curves shown in Figs. 3(c) and (d), we obtain

$$v_1 = f(i_1) = 2i_1 - \frac{1}{2}|i_1 + 1| + \frac{1}{2} \operatorname{sgn}(i_1 - 1) + \frac{1}{2}|i_1 - 2| \tag{31a}$$

$$i_2 = g(v_2) = -2 + \frac{3}{2} v_2 - \frac{1}{4}|v_2 + 2| + \operatorname{sgn}(v_2 + 2) + \frac{3}{4}\left(|v_2| - |v_2 - 1| + |v_2 - 3|\right) \tag{31b}$$

Since $g(\cdot)$ is strictly monotone-increasing, the composition $f \circ g(v_0)$ is well defined. Hence substituting the corresponding coefficients defining $f(i_1)$ and $g(v_2)$ from (31a) and (31b) into (29) and simplifying the resulting expression, we obtain

$$v_i \triangleq F(v_0) = -4 + 4v_0 - \frac{3}{4}|v_0+2| + 2\ \mathrm{sgn}(v_0+2) + \frac{3}{4}|v_0| + \frac{1}{2}\ \mathrm{sgn}(v_0 - \frac{1}{2}) - \frac{1}{2}|v_0-1| + \frac{3}{2}|v_0-3| \qquad (32)$$

Since the coefficients in (32) satisfy (17) with the strict inequality sign, it follows from Property 2 that the function F(•) is strictly monotone-increasing and therefore has a well-defined inverse $v_0 = F^{-1}(v_i)$, which is precisely the TC plot being sought. Hence, it remains to apply Property 4 to compute $F^{-1}(\cdot)$. Observe that since F(•) in (32) contains a discontinuity at $v_0 = \frac{1}{2}$, we must make use of (23) rather than (18) in computing for $F^{-1}(\cdot)$. After some routine calculation and simplification, we obtain:

$$v_0 = F^{-1}(v_i) = \frac{13}{15} + \frac{4}{15} v_i - \frac{1}{6}|v_i+7| + \frac{1}{3}|v_i+3| - \frac{1}{6}|v_i| - \frac{1}{6}|v_i - \frac{3}{2}| + \frac{1}{6}|v_i - \frac{5}{2}| + \frac{1}{12}|v_i-4| - \frac{3}{20}|v_i-8| \triangleq T(v_0) \qquad (33)$$

This TC plot is plotted as shown in Fig. 3(b) and can be easily verified by the graphical technique described in [5] to be the correct solution.

In addition to the many obvious applications to approximation and computation in device, circuit, and system modeling, the canonical piecewise-linear representation is also useful in carrying out many mathematical analysis of a circuit or system where the final results in explicit analytical form are desired. For example, this canonical form has been used to derive describing functions and Fourier coefficients in explicit form. A new existence and uniqueness theorem for a large class of resistive piecewise-linear networks has also been derived using the canonical form (1). Due to space limitation, we refer to reference 8 for details.

CONCLUDING REMARKS

The section-wise piecewise-linear representations presented in the preceding sections seem to be ideally suited for modeling nonlinear devices, circuits, and systems. The closed form canonical representations for f(x) are global in the sense that they are valid for all values of $\underset{\sim}{x} \in \mathbb{R}^n$. The coefficients characterizing these representations can be calculated efficiently and the amount of computer storage space is minimal in the sense that only the characterizing coefficients need be stored. Moreover, the "section-wise piecewise-linear" representations can allow finite jump discontinuities--a unique feature that should be useful for analyzing many circuit and control system containing devices characterized by discontinuous characteristics such as zener diodes, relay, etc.

REFERENCES

1. J. F. P. Hudson, Piecewise-Linear Topology, New York: W. A. Benjamin, 1969.
2. K. S. Cole, Membranes, Ions and Impulses, Berkeley: University of California Press, 1972.
3. A. A. Andronov, A. A. Vitt, and S. E. Khaikin, Theory of Oscillators, New York: Pergamon Press, 1966.
4. H. J. Zimmerman and S. J. Mason, Electronic Circuit Theory, New York: John Wiley, 1959.
5. L. O. Chua, Introduction to Nonlinear Network Theory, New York: McGran-Hill, 1969.
6. T. E. Stern, Theory of Nonlinear Networks and Systems, An Introduction, Reading, Mass.: Addison-Wesley, 1965.

7. K. K. Clarke and D. T. Hess, Communication Circuits: Analysis and Design, Reading, Mass.: Addison-Wesley, 1971.

8. L. O. Chua and S. M. Kang, "A New Analytical Representation for Section-wise Piecewise-Linear Functions and Its Applications," Berkeley, Calif: University of California, Electronics Research Laboratory, Memorandum ERL-M585, March 29, 1976.

9. L. O. Chua and P-M Lin, Computer-Aided Analysis of Electronic Circuits: Algorithms and Computational Techniques, Englewood Cliffs, N. J.: Prentice-Hall, 1975.

10. B. D. Wedlock and J. K. Roberge, Electronic Components and Measurements, Englewood Cliffs, N.J.: Prentice-Hall, 1969.

11. P. J. Davis, Interpolation and Approximation, New York: Blaisdell, 1963.

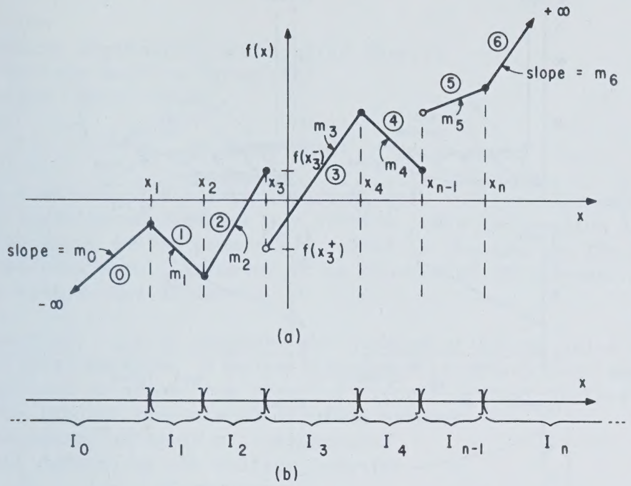


Fig. 1. A typical piecewise-linear curve with finite jump discontinuities

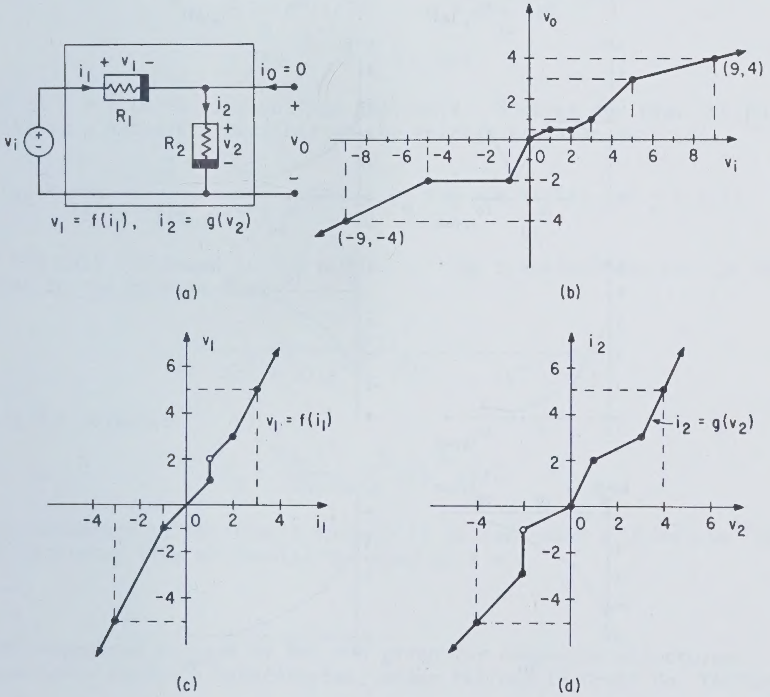

i_1
+ v_1 -
R_1
$i_0 = 0$
i_2
+
R_2
v_2
v_i
v_0
-
$v_1 = f(i_1)$, $i_2 = g(v_2)$
(a)
v_0
4
2
(9,4)
-8
-6
-4
-2
0
2
4
6
8
v_i
-2
-4
(-9,-4)
(b)
v_1
6
4
2
$v_1 = f(i_1)$
-4
-2
0
2
4
i_1
-2
-4
(c)
i_2
6
4
2
$i_2 = g(v_2)$
-4
0
2
4
6
v_2
-2
-4
(d)

Fig. 3. Derivation of TC plot of a nonlinear voltage divider

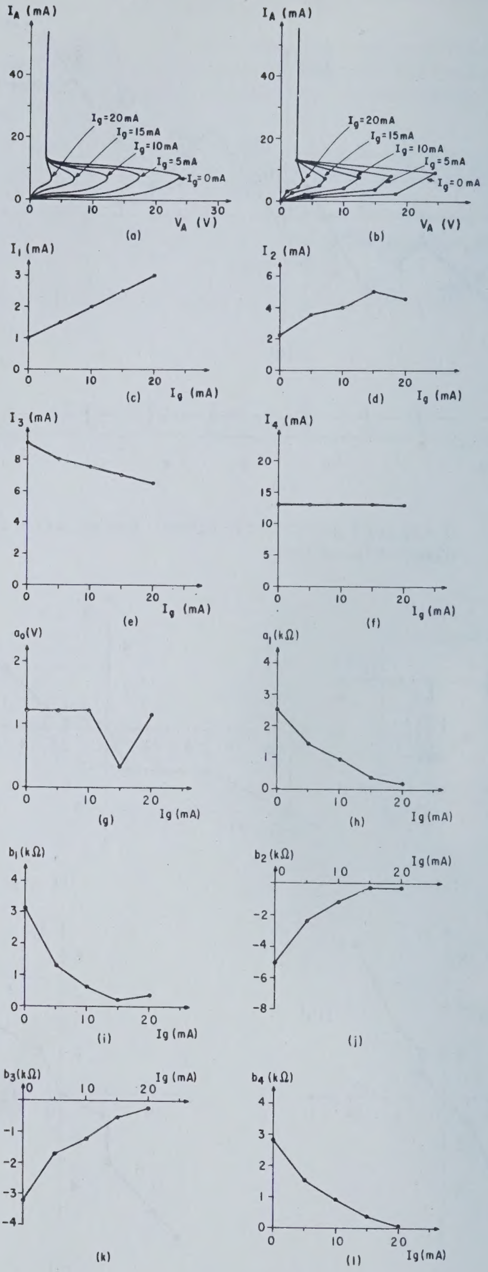


Fig. 2. The dc characteristics of a typical SCR and their piecewise-linear representations

SIGNAL PROCESSING VIA DISTRIBUTED FEED NETWORKS*

D. HAZONY
Electrical Engineering and Applied Physics
Case Western Reserve University
Cleveland, Ohio 44106

ABSTRACT

Pulse compression plays an important role in signal processing. It is shown that distributed feedback may provide pulse compression in a direct manner. The shape of the prescribed signal is the same as the shape of the required feedback path. Moreover it is shown that the process may be used to realize transversal filters.

1. Introduction. Let us consider the problem of square pulse compression by an electrical network. A pulse compressor separates overlapping signals. It might be used to select one group of signals against another. Also it uncovers the impulse response of a pulsed network.

Let the square pulse (Fig.1) be applied to a network defined by the voltage transfer function $t_{21} = (s^2 + 1)^{-1}$ where s is the Laplace Transform frequency variable. The output is:

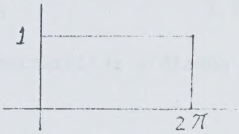


Fig.1

$$e_o = 2\sin^2(t/2) \qquad 0 \leq t \leq 2\pi \tag{1.1}$$

$$(=0 \qquad 0 \geq t \geq 2\pi)$$

Hence at $t = \pi$ there is a voltage gain of 2. Suppose now that the pulse is applied to a network prescribed by the voltage transfer function

$$t_{21} = \frac{2^2}{(s^2+1)(s^2+2^2)} \quad \text{then} \quad e_o = \frac{8}{3}\sin^4(t/2), \qquad 0 \leq t \leq 2\pi \tag{1.2}$$

Here the gain increased to 8/3 at $t = \pi$. The transfer function can be written in the general form:

$$t_{21} = \frac{(n!)^2}{(s^2+1)(s^2+2^2)\ \ldots\ (s^2+n^2)} \tag{1.3}$$

giving the response

$$e_o = \frac{(n!)^2}{(2n)!}\ 2^{2n}\sin^{2n}(t/2) \tag{1.4}$$

Taking advantage of Stirling's formula it is seen that $e_o \doteq \sqrt{n\pi}\,\sin^{2n}(t/2)$ which is rising like an impulse centered at $t = \pi$.

---

* Work supported in part by the NDE group for composite structures. Nasa Lewis Research Laboratories, under Fellowship Grant No. YAG7258.

The network described by 1.3 approximates the ideal square pulse compressor. It may be readily realized as a loss less network containing n reactive elements (see reference 1).

The idea of pulse shaping is discussed eloquently by Guillemin in the MIT Radiation series[2]. Recently it was demonstrated by J. D. Rhodes[3] that a pulse compressor can be approximated by cascaded transmission line segments. An interesting example is the Barker Code where a transversal filter compresses a sequence of square pulses[4]. In what follows the compressor as well as the transversal filter will be realized via distributed feedback and feedforward principles.

2. Distributed Feedback. Distributed feedback can be used to provide an ideal pulse to impulse compression in a direct manner. The concept is illustrated in Fig.2 where the emitter follower implies a unity gain amplifier. The resistive path provides a sufficiently low impedance so that essentially the voltage at each leg of the ladder is given by $E(x) = \alpha(x)E_2$ where

$$\alpha(x) = \int_0^x \rho(x)\,dx \Big/ \int_0^\lambda \rho(x)\,dx \qquad (2.1)$$

A possible realization is shown in Fig.3

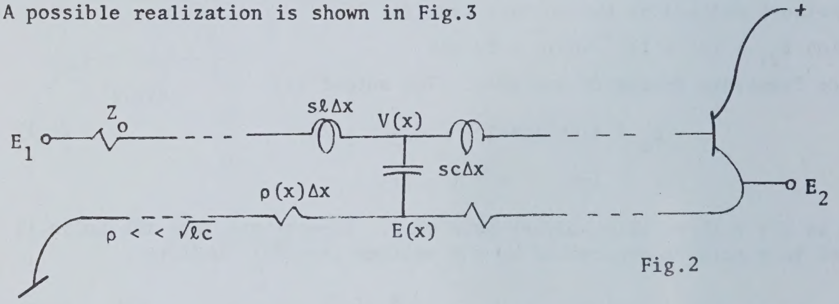


Fig.2

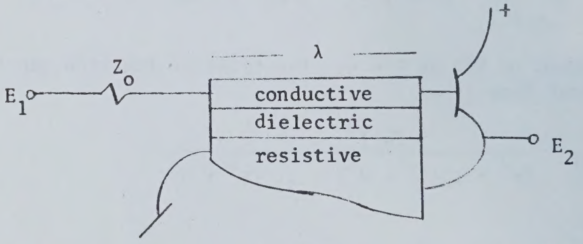


Fig.3

Without a loss in generality let $\int_0^\lambda \rho(x)\,dx = 1$ and let $\rho(x)$ define the taper of this network.

The concept of distributed feedback was proposed in reference 5 to realize the voltage transfer function $E_2 = E_1 e^{-s\tau}$ by an RC network.

3. Development. Considering the network of Fig.2 one may write

$$\frac{dV}{dx} + \ell s\, I = 0$$

$$\frac{dI}{dx} + sc\, [V - \alpha(x)\, E_2(s)] = 0 \qquad (3.1)$$

We proceed to apply the Laplace Transform in respect to x. Thus

$$U(p) = \int_o^\lambda V(x)e^{-px}dx, \quad W(p) = \int_o^\lambda I(x)e^{-px}dx, \quad \text{and} \qquad (3.2)$$

$$A = \int_o^\lambda \alpha(x)e^{-px}dx$$

and (3.1) becomes

$$pU - V(o) + \ell sW = 0 \qquad (3.3)$$

$$pW - I(o) + sc(U - AE_2) = 0$$

Solving for U and W gives: $(\tau = \sqrt{\ell c})$

$$U = \frac{pV(o) - s^2\tau^2 AE_2 - \ell sI(o)}{p^2 - s^2\tau^2} \qquad (3.4)$$

$$W = \frac{-scV(o) + scpAE_2 + pI(o)}{p^2 - s^2\tau^2}$$

Let ϕ define a new variable:

$$\phi = U + \sqrt{\ell/c}\;\; W = \frac{V(o) + s\tau AE_2 + \sqrt{\ell/c}\; I(o)}{p + s\tau} \qquad (3.5)$$

$$= \frac{V(o) + \sqrt{\ell/c}\; I(o)}{p + s\tau} + AE_2 - \frac{pAE_2}{p + s\tau}$$

which is readily inverted to give

$$E = L^{-1}\;\phi = [V(o) + \sqrt{\ell/c}\; I(o)]e^{-s\tau x} + \alpha(x)E_2 - E_2 \int_o^\lambda \rho(x - r)e^{-s\tau r}dr \qquad (3.6)$$

Here $\rho(x) = \frac{d\alpha}{dx}$ by definition.

The quantity $\vec{E}(x)$ of Eq.3.6 is a voltage function. Let the generator source impedance Z_o be equal to $\sqrt{\ell/c}$. Then

$$V(o) + Z_o I(o) = E_1 = \text{the generator voltage} \qquad (3.7)$$

Similarly, at the output of the network, (the base of the emitter follower), $I(\lambda) = 0$ and thus

$$\vec{E}(\lambda) = V(\lambda) + Z_o I(\lambda) = V(\lambda) = E_2 \qquad (3.8)$$

Moreover, $\alpha(\lambda) = 1$. It follows from Eq.3.6 that

$$E(\lambda) = E_2 = E_1 e^{-s\tau\lambda} + E_2 - E_2 \int_0^\lambda \rho(\lambda - r)e^{-s\tau r}dr \tag{3.9}$$

and

$$E_2 = \frac{e^{-s\tau\lambda}\ E_1}{\int_0^\lambda \rho(\lambda - r)e^{-s\tau r}dr} \tag{3.10}$$

Consider now an applied pulse whose frequency response is

$$E_1 = \int_0^T f(t)e^{-st}dt \tag{3.11}$$

The denominator of (3.10) will be cancelled if $r\tau = t$, $\tau\lambda = T$, and $\rho[(T - t)/\tau] = f(t)$. The network response would then be an impulse at $t = T$.

4. Special Cases. Suppose that the generator's impedance is zero in Fig.3, All the basic equation, 3.1 to 3.6, are essentially the same except that $\overrightarrow{E}(o) \neq E_1$. We rewrite Eq.3.6 and add to it $\overleftarrow{E}$

$$\overrightarrow{E} = [V(o) + \sqrt{\ell/c}\ I(o)]e^{-s\tau x} + \alpha(x)E_2 - E_2 \int_0^x \rho(x - r)e^{-s\tau r}dr \tag{4.1}$$

$$\overleftarrow{E} = [V(o) - \sqrt{\ell/c}\ I(o)]e^{s\tau x} + \alpha(x)E_2 - E_2 \int_0^x \rho(x - r)e^{s\tau r}dr$$

which is obtained from Eqs. 3.4 by subtraction instead of addition (see 3.5).

At $x = \lambda$, $\alpha(\lambda) = 1$, and $\overrightarrow{E} = \overleftarrow{E} = E_2$. Hence

$$V(o) + \sqrt{\ell/c}\ I(o) = e^{s\tau\lambda} \int_0^\lambda \rho(\lambda - r)e^{-s\tau r}drE_2 \tag{4.2}$$

$$V(o) - \sqrt{\ell/c}\ I(o) = e^{-s\tau\lambda} \int_0^\lambda \rho(\lambda - r)e^{s\tau r}drE_2$$

Adding these two equations and dividing by 2 gives:

$$V(o) = E_1 = \int_0^\lambda \rho(\lambda - r) \cosh s\tau(\lambda - r)drE_2 \tag{4.3}$$

Writing $\tau(\lambda - r) = t$ and $\lambda = T/\tau$ yields

$$E_2 = \frac{\tau\ E_1}{\int_0^\lambda \rho(t/\tau) \cosh stdt} \tag{4.4}$$

This is, therefore, a pulse compressor for a symmetric pulse. This arrangement is demonstrated in reference 6. Suppose now that Eqs.4.2 are subtracted. Hence

$$E_2 = \frac{\ell I_1}{\int_0^{\lambda} \rho(t/\tau) \sinh st\,dt} \tag{4.5}$$

which is a pulse compression for an asymmetric pulse.

The difference between these pulse compressors and the general case of section 3 is the total network delay. For the same pulse, the general compressor requires twice as much delay as the others. It might be said that in the symmetric and asymmetric case the first half of the signal excites the network and the second half relaxes it. In the general compressor, the generator resistor Z_o, provides an additional mechanism for relaxing the network.

5. Distributed Feedforward. Distributed feedforward can be used to obtain the transversal filter in a direct manner. The same distributed network will be employed but under a different electrical arrangement - Fig.4.

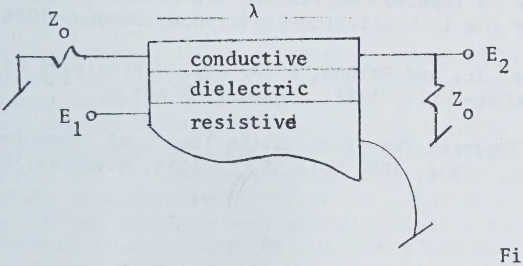


Fig.4

The basic equations are:

$$\frac{dv}{dx} + s\ell I = 0 \tag{5.1}$$

$$\frac{dI}{dx} + sc(V - \beta E_1) = 0 \; ; \quad \beta = 1 - \alpha$$

Following the development 3.1 - 3.6 gives

$$\vec{E} = [V(o) + Z_o I_o]e^{-s\tau x} + \beta(x)E_1 + \int_0^x \rho(x - r)e^{-s\tau r}dr E_1 \tag{5.2}$$

At $x = \lambda$, $E_2 = Z_o I_2$, $\vec{E} = 2E_2$, $\beta(\lambda) = 0$, and $V(o) + Z_o I(o) = 0$. Hence

$$E_2 = 1/2 \int_0^{\lambda} \rho(\lambda - r)e^{-s\tau r}dr E_1 \tag{5.3}$$

This is a voltage transfer function of the Distributed Transversal Filter.

A comprehensive discussion of transversal filters is given by Lucky, Salz, and Weldon in reference 7. According to them the device was invented by Wiener and Lee in 1935. It was first reported by Kallman in 1940. The transversal filter is usually approximated by tapped delay lines. Bolinder[8] (1950) used a similar concept in a truly distributed manner.

REFERENCES

1. Fialkow, A. D., and Gerst, I., "The Transfer Function of Networks Without Mutual Reactance," Quart, Appl. Math., Vol. 12, pp. 117-131, 1954.

2. White, H. J., Gillete, P. R., and Lebacqz, J. V., "The Pulse Forming Network," Radiation Laboratory Series No. 5 (Pulse Generators), McGraw-Hill, Chapter 6.

3. Rhodes, J. D., "Explicit Formula for Elements in Ideal Pulse Compression Transmission Line Networks", IEEE Transactions on Circuit Theory, Vol. CT-20, No. 2, pp. 150-153, March 1973.

4. Rihaczek, A. W., "Principles of High Resolution Radar," McGraw-Hill, 1969, Chapter 6.

5. Hazony, D., and Sobecks, R., "A Distributed RC Active Delay Line," SIAM J. Appl. Math. Vol. 22, No. 2, pp. 139-141, March 1972.

6. Hazony, D., "Pulse to Impulse Compression Via Distributed Feedback," The Proceedings of the 13th Allerton Conference, October 1975.

7. Lucky, R. W., Salz, J., and Weldon, E. J. Jr., "Principles of Data Communication," McGraw-Hill, 1968. Chapters 5 and 6.

8. Bolinder, E. F., "Fourier Transform in the Theory of Inhomogeneous Transmission Lines," Proc. IRE, Vol. 38, P. 1354, November 1950.

A NEWTON'S ALGORITHM FOR SOLVING SYSTEMS OF NONLINEAR EQUATIONS AND INEQUALITIES

H. MUKAI
Department of Systems Science and Mathematics
Washington University
St. Louis, Missouri 63130

Around 1970, Pshenichnyi [4] and Robinson [5] introduced extensions of Newton's methods for solving systems of equations and inequalities and established local and quadratic convergence for their methods. Since then several improvements of Robinson's method have been presented: [1], [2], and [3].

At each iteration all of these methods require the solution of a minimization subproblem whose constraints are the linearizations of some of the equations and inequalities around the current point. Unfortunately, all the methods except the one by Pshenichnyi disregard different degrees of relevance of each equation or inequality and utilize the linearizations of all the equations and inequalities as the constraints. Hence, where the subproblem may not have a solution the algorithm may reach an impasse even when that subproblem would have solutions once some of the less relevant constraints were deleted.

In this paper we present an algorithm, based on Pshenichnyi's idea, for solving systems of equations and inequalities and establish its global and quadratic convergence. At each iteration it requires a solution of a minimization subproblem as before. However, the constraints of this subproblem are the linearizations of a variable set of equations and/or inequalities. If the current set of equations and/or inequalities results in too high a dimensionality or in unsolvability of the subproblem then the algorithm deletes from the set the equations and/or inequalities that are least violated. Because of this feature the proposed algorithm is less likely to be thwarted by the unsolvability of the subproblem. Moreover, it has a better chance of coping with a large scale problem because of its capacity to delete some of the less relevant constraints and to reduce the dimensionality of the subproblem which is too large to handle otherwise.

[1] J.W. Daniel, "Newton's Method for Nonlinear Inequalities", Numer. Math. 21, 381-387, 1973.

[2] T.J. Huang, "Algorithms for Solving Systems of Equations and Inequalities with Applications in Nonlinear Programming", Computer Sciences Technical Report #191, Univ. of Wisconsin, 1973.

[3] E. Polak and I. Teodoru, "Newton Derived Methods for Nonlinear Equations and Inequalities", Nonlinear Programming 2, ed. by D.L. Mangasarian, R.R. Meyer and S.M. Robinson, Academic Press, 1975.

[4] B.N. Pshenichnyi, "Newton's Method for the Solution of Systems of Equalities and Inequalities", Math. Notes Acad. Sci. USSR, 8, 827-830, 1970.

[5] S.M. Robinson, "Extension of Newton's Method to Mixed Systems of Nonlinear Equations and Inequalities", Technical Summary Report No. 1161, Mathematics Research Center, Univ. of Wisconsin, 1971. Also, in revised form, "Extension of Newton's Method to Nonlinear Functions with Values in a Cone", Numer. Math. 19, 341-347, 1972.

PRIMITIVE CODING THEORY FOR CERTAIN CLASSES OF DISCRETE CHANNELS

William L. Root
The University of Michigan, Ann Arbor

ABSTRACT

The problem is considered of transmitting from a finite alphabet through a discrete channel which may consist of any one of a known finite class of mappings from input to output. The mapping which is effective during a certain time period is a priori unknown to transmitter and receiver, and may change from one time period to the next. These mappings are deterministic; there are no statistics involved. The maximum rate of guaranteed error free transmission is computed in some instances and bounded in others.

INTRODUCTION: THE CHANNEL MODEL

Certain kinds of unknown, finite-alphabet, discrete channels for which there are no statistical characterizations are described. The general problem which is posed is to find the greatest possible rate of transmission over such a channel with guaranteed error-free reception.

Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets with N_x and N_y elements, respectively. Denote elements of the product sets $\mathcal{X}^K$ and $\mathcal{Y}^K$ by $v = (x_1, \ldots, x_K)$, $x_i \in \mathcal{X}$, and $w = (y_1, \ldots, y_K)$, $y_i \in \mathcal{Y}$. Let Φ_K be a family of maps $\phi_1 : \mathcal{X}^K \to \mathcal{Y}^K$, $i = 1, \ldots, M$. $\mathcal{X}$ and $\mathcal{Y}$ are to be the sending and receiving alphabets of a communication channel which is deterministic, but unknown and possibly time-varying. We consider strings of input letters $\{x_i\}$ and corresponding strings of output letters $\{y_i\}$ partitioned into blocks of length K. Any one of these output blocks is given by $w = \phi_i(v)$ for one of the $\phi_i \in \Phi_K$; however in general the succeeding output block is given by $w' = \phi_j(v')$ where j is not necessarily equal to i. The class Φ_K is known to both sender and receiver, but the precise ϕ_i is not known, at least a priori, to either. For fixed K the class Φ_K is the same for each block.

A channel is a sequence of triples $\{\mathcal{X}, \mathcal{Y}, \Phi_K\}$, $K = 1, 2, \ldots$. A code of length K (or simply a code) is a subset of the N_x^K possible words of length K available to the sender for each block. A corresponding de-

coding rule assigns to each of the possible output words for one block (of number $\le N_y^K$) a code word. A code and decoding rule are *guaranteed error-free* if the decoding yields the correct code word no matter which code word was sent and no matter which ϕ_i is operative during the block.

Let $\Phi_K(v) = \{\phi_1(v), \ldots, \phi_M(v)\}$, the total image of a code word v. Two code words v_1 and v_2 are *distinguishable* if $\Phi_K(v_1)$ and $\Phi_K(v_2)$ are disjoint sets. Let $\mu = \mu(K) = \mu(\mathcal{X}, \mathcal{Y}, \Phi_K)$ be the largest number of distinguishable code words of length K for the channel $(\mathcal{X}, \mathcal{Y}, \Phi_K)$. Then the *K-rate* r_K for this channel is defined to be $r_K = \frac{1}{K} \log \mu(K)$.

After a few preliminaries in the next Section we consider in this paper only channels of a special type which are memoryless from symbol to symbol. Let Φ be a class of maps $\tilde{\phi}_i : \mathcal{X} \to \mathcal{Y}$. Then, for any K, the set of all maps ϕ_i of the form

$$\phi_i(v) = (\tilde{\phi}_{i_1}(x_1), \ldots, \tilde{\phi}_{i_K}(x_K)), \quad \tilde{\phi}_{i_j} \in \Phi$$

constitute a Φ_K. Now if for some fixed K the sequence of blocks can be arranged such that the same ϕ_i remains operative for each block, i.e. if $\phi_i(v) = (\tilde{\phi}_i(x_1), \ldots, \tilde{\phi}_i(x_K))$, we call the channel a *per-symbol channel of mesh K.* It is only this class of channels, with $K > 1$, which is dealt with in any detail.

A *time-invariant* per-symbol channel, is one for which the same ϕ_i remains operative indefinitely, i.e. it has arbitrary mesh K. For such channels the *asymptotic rate* r is defined to be

$$r = \lim_{K \to \infty} \frac{1}{K} \log \mu(K).$$

When discussing per-symbol channels we drop the tilde on the $\tilde{\phi}_i$ unless there is real danger of confusion. Thus we will write both $\phi_i(v)$ and $\phi_i(x)$.

PRELIMINARY RESULTS

The following facts are immediately evident.

(A) Let $\mathcal{X}, \mathcal{Y}$ and K be fixed and $\Phi_K \subset \Phi'_K$. Then $\mu(\Phi_K) \ge \mu(\Phi'_K)$.

(B) Let $\mathcal{X} \subset \mathcal{X}'$ $(N_x \le N_x')$ and let $\mathcal{Y}$ and K be fixed. Let Φ_K consist exactly of the restrictions to $\mathcal{X}$ of the maps $\phi' : \mathcal{X}' \to \mathcal{Y}$ in Φ'_K. Then $\mu(\mathcal{X}, \mathcal{Y}, \Phi_K) \le \mu(\mathcal{X}', \mathcal{Y}, \Phi'_K)$.

(C) $\mu(\mathcal{X}, \mathcal{Y}, \Phi_K) \le N_y^K$

(D) If Φ_K contains a map ϕ such that ϕ carries every element of $\mathcal{X}$ into the same element of $\mathcal{Y}$, then $\mu = 1$.

Consider now the case that $N_x = N_y = N$, and identify the elements of $\mathcal{X}$ with those of $\mathcal{Y}$ by an arbitrary bijection. With this identification of elements any map $\phi: \mathcal{X}^K \to \mathcal{Y}^K$ can be identified with a map $\mathcal{Y}^K \to \mathcal{X}^K$. In fact, we can think of the $\phi \in \Phi_K$ as being maps from $\mathcal{X}^K$ into itself; if ϕ is injective (and hence surjective) it has an inverse.

(E) If Φ_K is a group then the total images $\Phi_K(v)$ form the cells of a partition of $\mathcal{Y}^K$.

Proof: Suppose for some $\phi_1, \phi_2 \in \Phi_K$ and some v_1 and v_2 that $\phi_1(v_1) = \phi_2(v_2)$. Let ϕ_3 be any map in Φ_K. Then $\phi_3(v_1) = \phi_3\phi_1^{-1}\phi_2(v_2)$; that is, $\phi_3(v_1) = \phi_4(v_2)$ for some $\phi_4 \in \Phi_K$. Thus, $\Phi_K(v_1) = \Phi_K(v_2)$, and we see that the total images of elements in $\mathcal{X}^K$ either coincide or are disjoint. Since $\Phi_K(\mathcal{X}^K) = \mathcal{Y}^K$ the assertion follows. |||

(F) If Φ is a group of bijections from $\mathcal{X}$ to $\mathcal{Y}$ then any per-symbol channel formed from Φ has the property that the total images form the cells of a partition of $\mathcal{Y}^K$.

Proof: This follows from (E), since Φ_K is a group if Φ is. |||

(G) In the case that $N_x = N_y$ and Φ_K is a group, the relation $\sim$ between input words defined by: $v_1 \sim v_2$ if for some ϕ_1 and ϕ_2, $\phi_1(v_1) = \phi_2(v_2)$, is an equivalence relation.

Proof: This follows from (E). |||

In other words, the set of possible input words $v \in \mathcal{X}^K$ is also partitioned, and in fact the cells of this partition can be identified with the cells of the partition of $\mathcal{Y}^K$. Evidently a code consisting of one word from each equivalence class of $\mathcal{X}^K$, together with the decoding rule which assigns to each cell of the partition of $\mathcal{Y}^K$ the code word whose images lie in that cell, is guaranteed error-free.

(H) In the case that $N_x = N_y$ and Φ_K is a group, μ equals the number of cells in the partition of $\mathcal{X}^K$ (or $\mathcal{Y}^K$) determined by Φ_K.

Proof: Clearly μ is at least as large as this number. But if μ were larger, two code words v_i and v_j would have to lie in the same cell of the partition of $\mathcal{X}^K$, so that for some ϕ_1, ϕ_2, $\phi_1(v_i) = \phi_2(v_j)$. |||

This situation in which $\mathcal{Y}^K$ is partitioned by the total images of input words is the cleanest and simplest from a coding-theory point of view. We are of course also interested in situations in which this does not happen. From (A) and (H) one has the following simple bounds on μ in certain cases.

(I) If $N_x = N_y$ and each $\phi \epsilon \Phi_K$ is a bijection, then $\mu \geq$ the number of cells in the partition of $\mathcal{X}^K$ determined by the smallest group containing Φ_K, and $\mu \leq$ the number of cells in the partition determined by the smallest group contained in Φ_K.

PER-SYMBOL CHANNELS WITH $N_x = N_y$ AND Φ THE SYMMETRIC GROUP

Throughout this Section it is to be assumed that $N_x = N_y = N$ and that the channels are of the per-symbol type with mesh $K > 1$. We use letters a, b, c, ... to denote elements of $\mathcal{X}$ and $\mathcal{Y}$. A singular map $\phi : \mathcal{X}^K \rightarrow \mathcal{Y}^K$ or $\tilde{\phi}: \mathcal{X} \rightarrow \mathcal{Y}$ is a map that is not injective (and hence in this case, not surjective).

The fundamental example is that for which Φ is the group of all permutations on N letters. Then Φ_K consists of N! nonsingular maps, $\phi (v) = (\phi (x_1), \ldots, \phi (x_K))$, $\phi \epsilon \Phi$. (As previously indicated we also write $\phi (x) = \phi (x)$.) For convenience we refer to a channel with such a Φ as a full, nonsingular channel, and use the notation $\nu (N, K) \overset{d}{=} \mu (\mathcal{X}, \mathcal{Y}, \Phi_K)$.

(J) $\nu (N, K) = \sum_{i=1}^{\ell} S(i, K)$, where $\ell = \min (N, K)$, and the $S(i, K)$ are Stirling numbers of the second kind. We note that the Stirling number of the second kind, $S(i, K)$, is the number of ways the set of integers $\{1, 2, \ldots, K\}$ can be partitioned into i unordered classes (see, e.g., [1]).

Proof: Rather than fuss with a formal proof we look at an example; it will be clear how a general proof goes. Let K = 5 and N = 4. Put $\mathcal{X} = \{a, b, c, d\} = \mathcal{Y}$; consider the words v_1 = a b c d a, v_2 = c d a b c and v_3 = a a b c d. With $\phi_1 = \begin{pmatrix} a & b & c & d \\ a & b & c & d \end{pmatrix}$ and $\phi_2 = \begin{pmatrix} a & b & c & d \\ c & d & a & b \end{pmatrix}$ we have $\phi_1(v_1) = \phi_2 (v_2)$ = (a b c d a). Consequently $v_1 \sim v_2$. On the other hand, no matter what ϕ may be, $\phi (v_1)$ must have the same first and fifth letters with the other letters different, while any $\phi' (v_3)$ must have the same first and second letters with the others different. Hence v_1 is not $\sim v_3$. Obviously, $v \sim v_1$ if and only if the same letter appears in the first and fifth positions in v, with different letters in each other position. This situation can be described by saying v is equivalent to v_1 if and only if it has the pattern (15) (2) (3) (4). Similarly, a word will be equivalent to v_3 if and only if it has the pattern (12) (3) (4) (5). But (15) (2) (3) (4) and (12) (3) (4) (5) are partitions of the set $\{1, 2, 3, 4, 5\}$. In general, there is a 1:1 correspondence between the patterns of words of length 5 containing 4 letters and the partitions of $\{1, 2, 3, 4, 5\}$ into 4 classes. Each pattern represents one cell of the partition of $\mathcal{X}^5$ determined by Φ_5, so there are S(4, 5) cells of this partition which contain words with 4 distinct letters. We may pick one word of each pattern (one from each cell) to start the list of code words.

Now consider words using exactly 3 letters, e.g. v_4 = a b a c a, v_5 = b b c c d. v_4 has the pattern (135) (2) (4) and v_5 has the pattern (12) (34) (5). Clearly v_4 and v_5 are not equivalent, and neither v_4 nor v_5 nor any other word with 3 letters can be equivalent to any word with 4 letters. It follows as above that there are S(3, 5) patterns with 3 letters. We may pick S(3, 5) words, one with each of these patterns, to add to the code list.

Similarly, we can pick an additional S(2, 5) words with exactly 2 letters, and finally S(1, 5) = 1 code word with exactly 1 letter. We thus have one code word drawn from each of the cells of the partition of $\mathcal{X}^5$ determined by Φ_5. From (H) it follows that $\nu(4, 5) = S(1, 5) + S(2, 5) + S(3, 5) + S(4, 5)$. |||

Stirling's formula is

$$S(N, K) = \frac{1}{N!} \sum_{n=0}^{N} (-1)^{N-n} \binom{N}{n} n^K. \tag{1}$$

From this, or from recurrence formulas (see, e.g., [1]), one has for the example discussed above: S(1, 5) = 1, S(2, 5) = 15, S(3, 5) = 25, S(4, 5) = 10 and $\nu(4, 5) = 51$. It follows immediately that $\nu(5, 5) = 52 = \nu(N, 5)$ for $N > 5$.

(K) $\nu(N, K) \geq \frac{N^K}{N!}$ for all N and K. For fixed N, $\nu(N, K)$ is asymptotic to $\frac{N^K}{N!}$ as $K \rightarrow \infty$.

Proof: Using a simple argument involving the number of elements in the cells of the partition of $\mathcal{Y}^K$, one can verify the identity

$$\nu(N, K) = \frac{1}{N!} \left[N^K + S(N-2, K) \cdot N(N-1) \ldots 3(2!-1) + S(N-3, K) \cdot N(N-1) \ldots 4(3!-1) + \ldots + S(1, K) \cdot N((N-1)!-1) \right] \tag{2}$$

Hence $\nu(N, K) \geq \frac{N^K}{N!}$.

From Stirling's formula one sees that $S(N, K) \cdot \left(\frac{N^K}{N!} \right)^{-1} = 1 + o(1)$ as

$K \to \infty$. Then the conclusion follows from (J).

(L) A time-invariant full nonsingular channel has asymptotic rate $r = \log N$.

Proof: $\log \nu(N,K) = \log\left(\frac{N^K}{N!}\right) + \log\left(\nu(N,K) \cdot \frac{N!}{N^K}\right)$

$$= K \log N - \log N! + o(1). \quad |||$$

The following completely trivial observation seems interesting to the author. Suppose $\mathcal{C}$ is a full nonsingular channel with mesh $K \geq N$. Suppose a procedure is adopted in which first the particular channel mapping is identified, and from then on the sender is allowed to use the "full code" of all N-letter words—which the receiver can correctly decode because the map ϕ is known to him. It takes exactly (N-1) symbols to identify the channel completely. Thus the number of messages that can be transmitted with guaranteed error-free reception in K uses of the channel is $N^{K-(N-1)}$. But,

$$\nu(N,K) \geq \frac{N^K}{N!} \geq \frac{N^K}{N^{N-1}}$$

for any N and all K, with equality (for both inequalities) only when $N = 2$. Thus the two procedures are equivalent when $N = 2$; for $N > 2$, the universal code procedure always gives a larger K-rate than the procedure with identification first. If the channel is time-invariant, then the asymptotic rates are the same of course. This observation is stronger than perhaps it may first appear, because of two facts: one, for a time-varying channel the K-rate, when K is the mesh, is the largest attainable steady-state rate, so a difference in K-rates is meaningful; two, the comparison still holds even when Φ is not the full symmetric group if each $\phi \in \Phi$ is nonsingular and there is enough uncertainty about what maps actually are in Φ so that the communicators do not trust an identification made with fewer than (N-1) symbols.

SINGULAR PER-SYMBOL CHANNELS WITH $N_x = N_y$

A singular per-symbol channel is a per-symbol channel for which Φ contains at least one singular map. Once singular maps are permitted the problems of finding a biggest code or of computing r_K become very much more complicated than in the case of the preceding paragraph. The results presented here are rather weak and of a preliminary nature.

There is one simple conclusion that can be drawn immediately, viz.,

(M) With N fixed and $K \geq K(N)$ sufficiently large, any nonsingular channel is better than any singular channel, i.e. it has larger r_K.

Proof: Let μ_{ns} be the largest number of distinguishable K-code words for some (arbitrary) nonsingular channel $\mathcal{C}$ and μ_s be the corresponding number for some (arbitrary) singular channel $\mathcal{S}$. Since the Φ for any nonsingular channel is contained in the Φ for the full nonsingular channel, $\mu_s \geq \nu(N, K)$ by (A).

Let Φ' consist of any one singular map ϕ_1 which carries the N letters of $\mathcal{X}$ into N' symbols in $\mathcal{Y}$, where $N' \leq N-1$. Then, if $\mu' \overset{d}{=} \mu(\mathcal{X}, \mathcal{Y}, \Phi', K), \mu_1 \leq (N-1)^K$. It follows that

$$\frac{\nu(N, K)}{\mu'} \geq \left(\frac{N}{N-1}\right)^K \cdot \frac{1}{N!}$$

The singular channel $\mathcal{S}$ must have at least one singular map; hence, by (A), μ_s is no greater than some μ' as defined above. Therefore

$$\frac{\mu_{ns}}{\mu_s} \geq \frac{\nu(N, K)}{\mu'} \geq \left(\frac{N}{N-1}\right)^K \cdot \frac{1}{N!} \cdot |||$$

Now, for purposes of illustration consider two code words: $v_1 =$ (a b d a c a), $v_2 =$ (a b b a c d). The pattern of v_1 is (146) (2) (3) (5). This pattern corresponds to a partition of $\{1, 2, \dots, 6\}$ of type 3+1+1+1, or 1+3+1+1, etc., since the order does not matter (see [1] for a discussion of the combinatorics relating to the types of partitions); we will say simply the pattern is of type 3+1+1+1. v_2 has pattern (14) (23) (5) (6), of type 2+2+1+1. A nonsingular map ϕ preserves patterns; however a singular ϕ does not. For example if $\phi = \begin{pmatrix} a & b & c & d \\ d & d & a & b \end{pmatrix}$ then $\phi(v_1) =$ (d d b d a d), which has pattern (1246) (3) (5) of type 4+1+1. Let Φ be the set of all maps which carry 4 letters into 3 letters. Then the images under Φ of any v containing 4 letters will exhibit new patterns and types; there will in fact be $\binom{4}{2} = 6$ new patterns of various types. We illustrate for v_1 and v_2:

$$v_1\text{: } \Phi\,[(146)\,(2)\,(3)\,(5)] \longrightarrow \left[\begin{array}{l} (1246)\,(3)\,(5) \\ (1346)\,(2)\,(5) \\ (1456)\,(2)\,(5) \\ (146)\,(25)\,(3) \\ (146)\,(35)\,(2) \\ (146)\,(23)\,(5) \end{array}\right.$$

$$v_2\text{: } \Phi[(14)(23)(5)(6)] \longrightarrow \left[\begin{array}{l}(146)(23)(5)\\(145)(23)(6)\\(235)(14)(6)\\(236)(14)(5)\\(1234)(5)(6)\\(14)(23)(56)\end{array}\right.$$

With this Φ, the complete set of image patterns of v depends only on the pattern of v; for arbitrary Φ this is not true of course. Note that the sets of image patterns of v_1 and v_2 happen to intersect. We will call a pattern with m classes, an m-pattern. The Φ as prescribed carries each 4-pattern into a set of 3-patterns; furthermore each 3-pattern is an image of 4-patterns (for example, a 3+2+2 pattern is the image of five 4-patterns). When this Φ acts on a word with fewer than 4 letters, the situation is a little different. For example,

$$\Phi[(1234)(5)(6)] \longrightarrow \left[\begin{array}{l}(1234)(5)(6)\\(12345)(6)\\(12346)(5)\\(1234)(56)\end{array}\right.$$

The original pattern is repeated in the image set because some of the maps in Φ carry the 3 letters of v into 3 distinct letters.

This discussion generalizes in an obvious way to arbitrary N and K. Let Φ_p denote the class of all maps from N letters into N-p letters, p = 0, 1, . . . , N-1. Then the discussion also generalizes directly to classes Φ which are a Φ_p or a union of Φ_p's. It does not generalize to arbitrary classes Φ because the image patterns for a word v will in general depend on v and not just on the pattern of v. To keep things simple we consider here only channels for which $\Phi = \Phi_1$ or $\Phi = \Phi_o \cup \Phi_1$.

An (N, K) - graph is described as follows. On one horizontal line list all N-patterns for words of length K. On a horizontal line below this list all (N-1)-patterns for words of length K. On a horizontal line below this, all (N-2)-patterns. Keep this up until the last line has the single entry (1, 2, . . . , K). These entries are the vertices of the graph. From each vertex (N-pattern) in the top line draw a line to each vertex ((N-1)-pattern) in the second line which is an image of this N-pattern under Φ_1. Similarly, connect vertices in the second line ((N-1) -patterns) to their images in the third line under Φ_1. The only image of an (N-1)-pattern that is an (N-1)-pattern is itself, so there are no horizontal connecting lines. Keep up this procedure as far as possible; finally all vertices in the next-to-bottom line will connect to the single vertex at the bottom.

The N-patterns, (N-1)-patterns, . . . , 1-pattern of length K are a partially ordered set under the ordering given by this directed graph, but they are not a lattice. A *1-tree* is a vertex and all the vertices of the line below to which it is connected; the vertex on the higher line is the *base vertex* for the tree.

(N) Let $\mathcal{C}$ be a per-symbol channel of mesh K with $N_x = N_y = N$ and $\Phi = \Phi_1$. Then μ is equal to the largest number of disjoint 1-trees that can be imbedded in the (N-K)-graph. A biggest code is given by any set of code words whose patterns are those of the base vertices in a maximum imbedding of disjoint 1-trees.

Proof: It is clear that any code for which the code words have patterns corresponding to the base vertices of disjoint 1-trees is error-free, because each code word has images with distinct patterns and hence is distinguishable from the others. Conversely, any error-free code must have the property that the 1-trees corresponding to the code words must be disjoint. This is because of the fact that if two code words have a common image pattern, they have common image words. Since there are only a finite number of imbeddings with disjoint 1-trees, there are such imbeddings with a maximum number of 1-trees. The conclusions follow. |||

(O) Let the μ of (N) be denoted μ_1. With the same situation as in (N) except that $\Phi = \Phi_0 \cup \Phi_1$ denote the μ by μ_{01}. Then $\mu_{01} = \mu_1$.

Proof: $\mu_{01} \leq \mu_1$ by (A). However any error-free code with Φ_1 is also an error-free code with $\Phi_0 \cup \Phi_1$ since the nonsingular maps leave all patterns invariant. |||

For channels of the type described in (N) and (O) an upper bound on μ is given by:

(P) $$\mu \leq 2 \left(\frac{S(N-1, K)}{N(N-1)} + \frac{S(N-2, K)}{(N-1)(N-2)} + \ldots + \frac{S(1, K)}{2 \cdot 1} \right) \qquad (3).$$

This bound is not attainable.

Proof: There are S(N-1, K) vertices at the (N-1) level of the (N-K)-graph. Any code word with N letters will have $\frac{N(N-1)}{2}$ image patterns at the (N-1) level. Hence, even with perfect packing there cannot be more than $\frac{2 \cdot S(N-1,K)}{N(N-1)}$ disjoint 1-trees with base vertex at the N level. The argument is repeated at lower levels. The bound is not attainable, because if there were perfect packing at any level there could be no base vertices at that level. |||

This bound is admittedly somewhat crude, partly because it takes no account of pattern types. The packing problem is not simple because a pattern of a given type has images with a certain distribution of types, and these distributions are different. However, there are symmetries to exploint, which the author has not yet done.

Maximal error-free codes, which are not necessarily biggest error-free codes of course, can be constructed by starting at some level (not necessarily N) and systematically eliminating vertices of disjoint 1-trees. This procedure is simple and straightforward.

Example. N = 4, K = 5. The following sets of patterns each give biggest error-free codes for this case:

(i) (12) (34) (5), (13) (45) (2), (14) (25) (3)
(15) (23) (4), (24) (35) (1); (12345)

(ii) (13) (2) (4) (5); (12) (34) (5), (14) (25) (3),
(15) (23) (4), (24) (35) (1); (12345).

The discussion of this Section extends in a fairly obvious way to any case where Φ is a union of Φ_p's.

ACKNOWLEDGMENT

This research is sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under AFOSR Grant Number 72-2328.

REFERENCE

[1] C. Berge, Principles of Combinatorics, Academic Press, N. Y.(1970).

ON TRUNCATED VITERBI DECODING

DANIEL J. COSTELLO, JR. AND FARHAD HEMMATI
Department of Electrical Engineering, Illinois Institute of Technology
Chicago, Illinois 60616

SUMMARY

In Viterbi decoding of convolutional codes, the best path leading to each state must be stored. For long data sequences, this leads to excessive storage requirements, and it is common practice to truncate the path memory and to make final decoding decisions with a fixed decoding delay. We will assume that a final decoding decision is made on the branch at time t by choosing the path with the best metric at time $t + \tau$ and decoding the information bit at time t on that path. The decoding delay of τ branches is called the truncation length of the decoder. A truncation error is said to occur at time t if an incorrect branch is decoded which would not be chosen by a Viterbi decoder without truncation.

In this paper we bound the truncation error probability for a specific code in terms of the distance properties of the code. A truncation error occurs if a path which diverges from the correct path at time t, and stays unmerged with the correct path through time $t + \tau$, is the best path at time $t + \tau$. The truncation error probability P_T is therefore upper bounded by the probability that any such path has a better metric than the correct path at time $t + \tau$. Assuming the all-zero path to be the correct path, the generating function for the indicated set of paths can be obtained by summing the transfer functions connecting the all-zero state to every other state in the state diagram of the encoder, and retaining only those terms corresponding to paths of length τ. Letting $T'(D, L^\tau)$ be the desired generating function, and following Viterbi's[1] union bound approach, we can bound the truncation error probability for the binary symmetric channel as

$$P_T < T'(D, L^\tau)\Big|_{D = 2\sqrt{p(1-p)}}, \qquad (1)$$

where p is the crossover probability of the channel.

Let d_τ be the smallest power of D in the bound of (1). If $d_\tau > d_{free}$, the free distance of the code, then the truncation error probability will be small compared to the error probability without truncation. On the other hand, if $d_\tau \leqq d_{free}$, truncation will add significantly to the overall decoding error probability. Hence, for a specific code, the truncation length τ should be chosen large enough so that the truncation distance d_τ exceeds the free distance d_{free} of the code. In other words, the minimum weight of all unmerged codewords of length τ should exceed d_{free}.

ACKNOWLEDGEMENT

This work was supported by the National Science Foundation, under Grant ENG 75-20068, at the Illinois Institute of Technology, Chicago, Illinois 60616.

REFERENCES

1. Viterbi, A. J., "Convolutional Codes and Their Performance in Communication Systems", IEEE Trans. Comm. Tech., Vol. COM-19, Part II, pp 751-772, October 1971.

REDUCED-STATE VITERBI DECODING OF CONVOLUTIONAL CODES

JEAN CONAN
Département de Génie Electrique
Ecole Polytechnique
Montréal, Canada

DAVID HACCOUN
Département de Génie Electrique
Ecole Polytechnique
Montréal, Canada

ABSTRACT

This paper presents decoding techniques for convolutional codes which use the trellis structure of these codes and yet attempt to reduce the computational load of Viterbi decoding. For a binary rate 1/n convolutional code of memory order m, a reduced-state trellis is first defined where the number of states may be 2^{m-1}, 2^{m-2}, ..., etc. A reduced-state Viterbi decoder is then defined as a maximum likelihood decoder operating on the reduced-state trellis. As a consequence the number of computations for every decoded bit is 2^{m-1}, 2^{m-2}, ..., etc, [as compared to 2^m with straight Viterbi decoding]. Such a decoder is suboptimal and moreover can generate catastrophic error events which make the procedure inapplicable to real time bit decoding. A new class of generalized reduced-state Viterbi decoders which alleviate this difficulty is presented. These procedures adapt the computational load to the current noise level and can operate at substantially higher speed than straight Viterbi decoders without degradation in error performance.

I. INTRODUCTION

Two of the most powerful techniques for decoding convolutional codes on Discrete Memoryless Channels (DMC) are sequential decoding and Viterbi decoding. Both schemes attempt to find the most likely path through a graph in which the branches are assigned "branch metrics" or likelihood values. A Viterbi decoder systematically explores at each step all possible paths that may have been transmitted, discarding some of the incorrect paths. For a convolutional code of rate 1/n and memory order m, the algorithm always performs 2^m computations per decoded bit. Viterbi decoding is thus limited to short memory codes ($m \leq 7$). By contrast the computational effort of sequential decoding is variable (with an asymptotic Pareto distribution), but on the average typically very small. An improvement of both the error performance and the computational distribution can be achieved by the larger class of generalized sequential decoders [1], but asymptotically, the Pareto distribution appears to be inescapable. This paper presents some decoding alternatives which eliminate the Pareto behavior of the computational distribution by never allowing the decoder to backtrack and revise earlier decisions. We assume the reader to be familiar with the tree and trellis structure of convolutional codes. Without loss of generality, we only consider binary codes of rate 1/n and constraint length $K = m+1$, where m is the memory order of the code (Fig. 1). In section II Viterbi decoding is quickly reviewed and the reduced-state Viterbi decoder [2] is presented in section III, emphasizing the catastrophic error propagation. A new class of generalized Viterbi decoders with a variable number of computations per decoded bit is presented in Section IV together with some computer simulation results.

It is shown that although variable, the distribution of the number of computations per decoded bit is upperbounded by that of Viterbi decoding but with an average value that may be substantially smaller.

II. VITERBI DECODING

II-1 Convolutional Encoding

Let $\underline{U} = u_0, u_1, u_2, \ldots$ be the input information sequence to the encoder. Then the corresponding output sequence is $\underline{X}^{\underline{U}} = \underline{x}_0, \underline{x}_1, \underline{x}_2, \ldots$ where $\underline{x}_t = (x_t^1, x_t^2, \ldots, x_t^n)$ is the branch output corresponding to the information bit u_t at time t. Denoting the n code generators by

$$g_j = (g_j^0, g_j^1, \ldots, g_j^m), \quad 1 \le j \le n \tag{1}$$

we have

$$\underline{x}_t^i = \sum_{k=0}^{m} u_{t-k}\, g_i^k, \quad 1 \le i \le n \tag{2}$$

The state s_t of the convolutional encoder at time t is the contents of the m shift register stages at time t, and the state space S is the set of all 2^m possible states: $S = \{s_t = (u_{t-1}, u_{t-2}, \ldots u_{t-m})\}$. The encoder state together with the next input bit uniquely specifies the n output symbols. Fig. 2 shows the state diagram (or Good-De-Bruiyn graph) for the rate $\frac{1}{2}$, m = 1 encoder with generators $g_1 = (10)$ and $g_2 = (11)$.

At time t = 0, the encoder is usually assumed to be in the "all zero" state. The trellis diagram of the convolutional encoder is obtained by linking all possible state values at time t to all possible state values at time (t + 1) by the corresponding coded branches. The trellis diagram has 2^m states and may be terminated at some time L + m by appending a tail of m 0's to the L information bits. Whenever L goes to ∞, the trellis is unterminated. For example, part of the unterminated trellis for the code of Fig. 2 is shown in Fig. 3.

II.2 Viterbi Decoding

Let $\underline{Y}$ be the received sequence when $\underline{X}^{\underline{U}}$ is transmitted over a DMC. Decoding is then equivalent to finding the path $\tilde{\underline{U}}$ in the trellis such that $\underline{X}^{\tilde{\underline{U}}}$ is "closest" to $\underline{Y}$. An error will occur whenever $\tilde{\underline{U}} \ne \underline{U}$. The sequence error probability P_E is minimized by applying the maximum likelihood decoding rule: observe $\underline{Y}$ and choose $\tilde{\underline{U}} = \underline{U}^*$ if

$$\log P(\underline{Y}|\underline{X}^{\underline{U}^*}) \ge \log P(\underline{Y}|\underline{X}^{\underline{U}}), \quad \forall\, \underline{U}^* \ne \underline{U} \tag{3}$$

Let $\gamma_i^{\underline{U}} = \log P(\underline{Y}|\underline{x}_i^{\underline{U}})$ be the log-likelihood or branch "metric" for the i^{th} branch of path $\underline{U}$. Then, for a DMC, the log-likelihood or metric of path $\underline{U}$ over the first N branches is

$$\Gamma_N^{\underline{U}} = \sum_{i=1}^{N} \gamma_i^{\underline{U}} \tag{4}$$

For a sequence of length L branches, the decoded rule (3) is then: observe $\underline{Y}$ and find $\underline{\tilde{U}}$ such that

$$\Gamma^{\underline{\tilde{U}}}_{L+m} = \max_{\underline{U}'} \Gamma^{\underline{U}'}_{L+m} \tag{5}$$

In implementing this decoding rule, a Maximum Likelihood Viterbi Decoder (MLVD) calculates the metric of each path entering a given state and retains as survivor at that state the one which is the most likely. This is done for each of the 2^m states at a given trellis depth $\ell > m$. The procedure is repeated for each level of the trellis until the end is reached. The decoded sequence is then the input sequence which retraces in the trellis the surviving path at state zero.

Defining a computation as being the sequence of operations which lead to the extension of one state in the trellis, the amount of computation required to decode an information sequence of L bits is

$$C \cong (L + m)2^m \tag{6}$$

In applications requiring high efficiency, the sequence length L must be made large with regard to m. On the other hand, as L is increased, it must be the case that P_E tends to 1 for any practical channel. In this case, as well as in situations where the bit and not the sequence is the pertinent information, the criterion of interest is the bit error probability

$$P_{BE} \triangleq \frac{1}{L} \sum_{\ell=1}^{L} P(\tilde{u}_\ell \neq u_\ell) \tag{7}$$

which gives the fraction of bits which are wrongly decoded. Algorithms which minimizes P_{BE} have been proposed both for the terminated trellis [3] and the unterminated trellis case [4]. It appears essentially that for real time applications, these algorithms do not provide any substantial gain over a simple suboptimal modification of the MLVD. Such an algorithm, termed Real Time Viterbi Decoder (RTVD) can be outlined as follows. Its operation is identical to that of the MLVD except that the recursive sequence of extensions and eliminations is repeated indefinitely. As such the total number of computations per decoded bit is

$$C = 2^m \tag{8}$$

Bit decoding is performed as follows. Given a decoding lag value $\Delta > m$, whenever $\ell \geq \Delta$, the decoded bit is the symbol at depth $\ell-\Delta$ on the surviving path which is the most likely among all survivors at step ℓ. An alternative practical decoding rule is to select the decoded bit on a majority basis among the symbols at depth $\ell-\Delta$ on all the survivors at step ℓ.

III. REDUCED-STATE VITERBI DECODING

It was seen in the preceding section that no matter which version of the Viterbi decoder is used, the amount of computation per decoded bit grows exponentially with the code memory order m. In order to reduce the computational load, the state space dimensionality must be somehow reduced. Such a reduction can be obtained by defining equivalent classes in the state space.

III.1 Reduced-State Viterbi Decoding Algorithm

Given any positive or null integer $\mu \leq m$ (μ is the decoder memory order), we first define an equivalence relation on the state space S as follows.
Definition: Two states s^1, s^2, $\in S$ are said to be R_μ- equivalent if the two binary m-tuples associated with these two states are identical in their rightmost μ symbols.
Definition: The μ- reduced state space S_μ is defined as the quotient of S by the equivalence relation R_μ.

As a consequence, S_μ contains 2^μ distinct "pseudo" states and each "pseudo" state is equivalent to exactly $2^{m-\mu}$ real states. A μ-reduced state trellis (terminated or unterminated), containing 2^μ "pseudo states" at each depth is then obtained by reducing the original trellis through the R_μ equivalence relation. A μ-reduced state Viterbi decoder is then a Viterbi decoder operating on the μ-reduced trellis as follows.

The μ-reduced state MLVD algorithm [For information sequence of length L]

```
Step 1: Extend* in the $\mu$-reduced trellis all distinct $2^\mu$ paths diverging
        from the root node $s_0^\mu$ over $\mu$ branches and compute their metrics.
        Call these paths the "pseudo survivors". Set $\ell = \mu$
Step 2: Extend* all the $2^\mu$ paths surviving at depth $\ell$ and compute their metrics.
        This results in two paths which are "pseudo reconvergent" at each one of the
        "pseudo states" of the reduced trellis at depth $\ell + 1$. For each node at
        depth $\ell + 1$, compare the metrics of the two "pseudo reconverging" paths
        and retain as "pseudo survivor" the one with the largest metric (in case of
        ties pick one at random).
Set $\ell = \ell + 1$
    If $\ell = L + \mu$, stop (the decoded path is the "pseudo survivor" at node 0)
    Else, go to step 2.
```

This algorithm can also be modified for real time applications by using the procedure discussed in section II. It is readily seen that both algorithms perform roughly 2^μ computations per decoded bit, and when $m = \mu$ they are equivalent to straight Viterbi decoding.

As a sequence decoder, the reduced state Viterbi algorithm is obviously suboptimal whenever $\mu < m$. If $P_E^{\mu,m}$ is the sequence error probability for a μ-reduced state Viterbi decoder operating on a memory m convolutional code then

$$P_E^{\mu,m} \geq P_E^{m,m} \qquad \mu \leq m \tag{9}$$

---

* Extension of a path from a "pseudo state" is always performed by using the real state value associated with this path before extension.

III.2 Catastrophic Error Propagation

As a real time decoder, a reduced-state Viterbi decoder is subject to catastrophic error propagation, in the sense that a finite number of channel errors can result in an infinite number of decoded errors. It is well known that a Viterbi decoder is immune to catastrophic error propagation provided the code satisfies the Massey and Sain condition [5]. In the case of a reduced state Viterbi decoder, the analysis of error catastrophe is far more complicated. To illustrate, let us consider the simple code of Fig. 2, and let us assume a BSC with crossover probability p. The code of Fig. 2 satisfies the Massey and Sain and therefore does not give error propagation with straight Viterbi decoding. Assuming $\mu = 0$ and that ties are resolved at random, the dynamical behavior of the decoder can be described by a two-state stochastic sequential machine, whose transition diagram is represented in Fig. 4. Assuming that zeroes are transmitted, whenever state $\ulcorner 1 \urcorner$ is reached, a bit error is made. Following this error, and even if no subsequent channel error occurs, the probability of a solid burst of b subsequent decoded errors is $(\frac{1}{2})^b$. Hence there is a non zero probability that errors will propagate over any arbitrary length. In the simple case of Fig. 2, steady state probalities can be computed for the states $\ulcorner 0 \urcorner$ and $\ulcorner 1 \urcorner$

$$P\{\ulcorner 0 \urcorner\} = \frac{1}{1+2p} \qquad\qquad (10)$$

$$P\{\ulcorner 1 \urcorner\} = \frac{2p}{1+2p}$$

Since state $\ulcorner 1 \urcorner$ implies a decoded error and reciprocally, the bit error probability for this example is simply

$$P_{BE} = P\{\ulcorner 1 \urcorner\} = \frac{2p}{1+2p} > p, \quad 0 < p < \tfrac{1}{2} \qquad (11)$$

As a consequence, the performance of this decoder is worse than the uncoded system over all the possible range of operation. This illustrates the fact that reduction of the trellis dimensionality may lead to serious problems when used in connection with real time bit decoding.

IV. ADAPTIVE REDUCED-STATE VITERBI DECODING

As seen in the preceding section, the main drawback of reduced state Viterbi decoders is the possibility of error catastrophic propagation. This is due essentially to the fact that the path elimination procedure is too rigid. In fact it may be the case that a path terminating in a given equivalence class will be eliminated even if its metric is larger than that of a "pseudo survivor" in a different equivalence class. On the other hand, for any code, a Viterbi decoder performs a large number of computations even when the noise level is low. This suggests the use of an adaptive scheme which performs as a Viterbi decoder when the channel is noisy and as a reduced-state Viterbi decoder when the channel is quiet. Such a decoding algorithm can be outlined as follows:

Given $\mu < m$, a finite list with a maximum of $2^{m-\mu}$ entries is assigned to each of the "pseudo states" in S_μ. The algorithm operates as a Viterbi decoder, extending all paths in the lists and taking care of reconvergence.

At the end of the extension cycle, a purge is performed on the survivors as follows. Given a Threshold value $T \geq 0$, for every list i the largest path metric in the list Γ_i^* is computed and all the active paths in this list with metric less than $\Gamma_i^* - T$ are eliminated. At the start of the procedure, all the lists are empty except the list which contains the root node.

The class of algorithms so defined is quite large depending on the values chosen for the threshold T and the code memory μ. We note that if $\mu = m$, the purging cycle becomes inoperative and the algorithm becomes a true Viterbi decoder. When $T = 0$, the algorithm operates as a reduced-state V.D. with unresolved ties between "pseudo reconvergent" paths. The purging cycle allows the procedure to eliminate quickly paths which look very unlikely. Whenever a bad burst of noise occurs, almost all paths will seem to appear good. As a consequence, if the threshold value is not too tight, the procedure will behave locally as a true Viterbi decoder until the end of the noise pattern, at which time a smaller number of paths will be followed depending on the values of μ and T.

Computer simulations of this new decoding procedure have been performed in the case of a BSC with $p = .05$ and for the rate $\frac{1}{2}$, $m = 4$, $g_1 = (11101)$, $g_2 = (10011)$ code [6]. To investigate error propagation, the algorithm was programmed as a bit decoder with a decoding delay of 36 bits. All runs implied 95964 decoded bits. The results of the simulation are summarized in Table 1 for different values of μ and T. Typical distributions for the number of computations per decoded bit are given in Fig. 5. It appears that the properties of these adaptive algorithms can be summarized as follows:

1) Error catastrophic propagation can be avoided without requiring all the time the full extension of the trellis.
2) The performance can be very close to that of Viterbi decoding but with a substantially smaller computational load.
3) The number of computations per decoded bit is a random variable with an average value $2^\mu \leq \bar{C} \leq 2^m$, and a maximum value $C_{max} \leq 2^m$

V. CONCLUSION

A new class of generalized Viterbi decoders has been presented. Given μ, the decoder memory, a wide range of operations can be obtained by variation of a purging threshold T. In general the average number of computations per decoded bit can be made smaller than that of straight Viterbi decoding, and at almost no expense on the decoding performance. The variability of the computational load is the price to be paid. However, in contrast with sequential decoding, the maximum number of computation per decoded bit remains bounded. As a consequence, for a decoder which takes advantage of the speed factor of the algorithm, the probability of overflow of the input buffer can be easily controlled leading to an erasure-free operation.

REFERENCES

[1] D. Haccoun, M.J. Ferguson, "Generalized Stack Algorithms for Decoding Convolutional Codes", IEEE Transactions on Information Theory, Vol. IT-21, No. 6, November 1975.

[2] F. Jelinek, "Bootstrap Trellis Decoding", IEEE Transations on Information Theory, Vol. IT-21, No. 3, May 1975.

[3] L.R. Bahl, J. Cocke, F. Jelinek, and J. Raviv., "Optimal Decoding of Linear Codes for Minimizing Symbol Error Rate", presented at the 1972 Int. Symp. Information Theory, Asilomar, Calif., January 1972.

[4] L.N. Lee, "Real-Time Minimal-Bit-Error Probability Decoding of Convolutional Codes", IEEE Transactions on Communications, Vol. COM-22, No. 2, February 1974.

[5] J.L. Massey, M.K. Sain, "Inverses of Linear Sequential Circuits", IEEE Transactions on Computers, Vol. C-17, April 1968.

[6] J.P. Odenwalder, "Optimal Decoding of Convolutional Codes", Ph.D. Dissertation, Department of Electrical Engineering, U.C.L.A., January 1970.

μ	T	# of errors	P_{BE}	$\bar{C}$	Behavior of the algorithm
0	0	48412	$\simeq$ .5	1.34	Error propagation
	1	12267	.128	2.73	"
	2	647	$6.74\ 10^{-3}$	4.88	Adaptive
	3	551	$5.74\ 10^{-3}$	11.15	"
	4	515	$5.37\ 10^{-3}$	14.23	"
	5	503	$5.24\ 10^{-3}$	$\simeq$ 16.	Viterbi
1	0	48412	$\simeq$ .5	2.	Error propagation
	1	2141	$2.23\ 10^{-2}$	4.83	"
	2	575	$5.99\ 10^{-3}$	6.64	Adaptive
	3	531	$5.53\ 10^{-3}$	13.43	"
	4	499	$5.2\ 10^{-3}$	14.97	"
	5	503	$5.24\ 10^{-3}$	$\simeq$ 16.	Viterbi

Table 1. Results of the simulation runs. BSC, $p = .05$. Code: $R = \frac{1}{2}$, $m = 4$, $g_1 = (11101)$, $g_2 = (10011)$

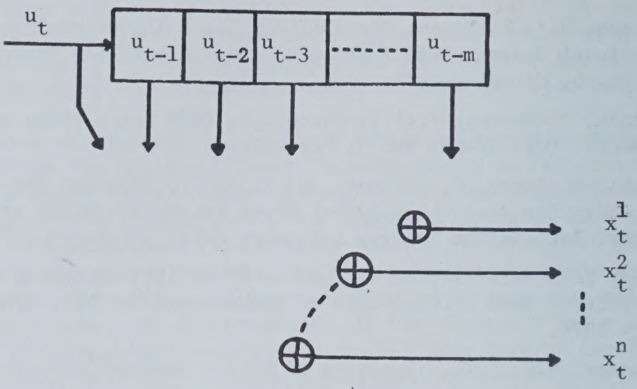

u_t
u_{t-1}
u_{t-2}
u_{t-3}
u_{t-m}
x_t^1
x_t^2
x_t^n

Figure 1. General rate 1/n convolutional encoder

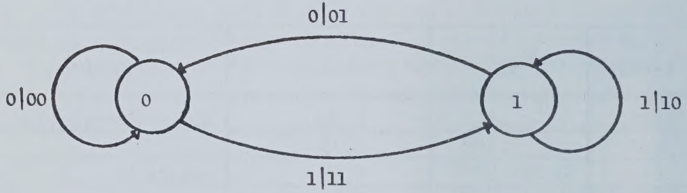

0|01
0|00
0
1
1|10
1|11

Figure 2. State transition diagram for the rate $\frac{1}{2}$, m=1, $g_1 = (10)$, $g_2 = (11)$ convolutional code

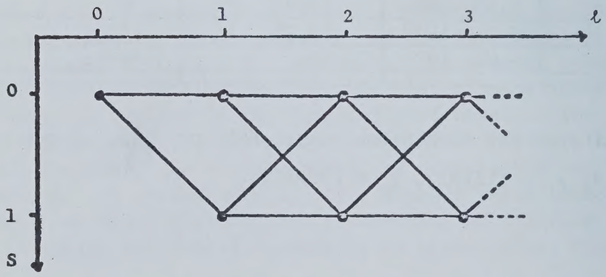


Figure 3. Unterminated treillis for a rate $\frac{1}{2}$, m = 1 convolutional code

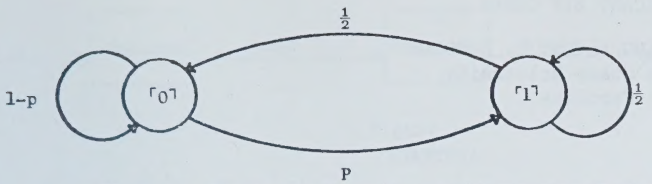


Figure 4. State transition diagram for the two-state stochastic sequential machine equivalent to the 0-reduced-state V.D. of the convolutional code of Fig. 2.

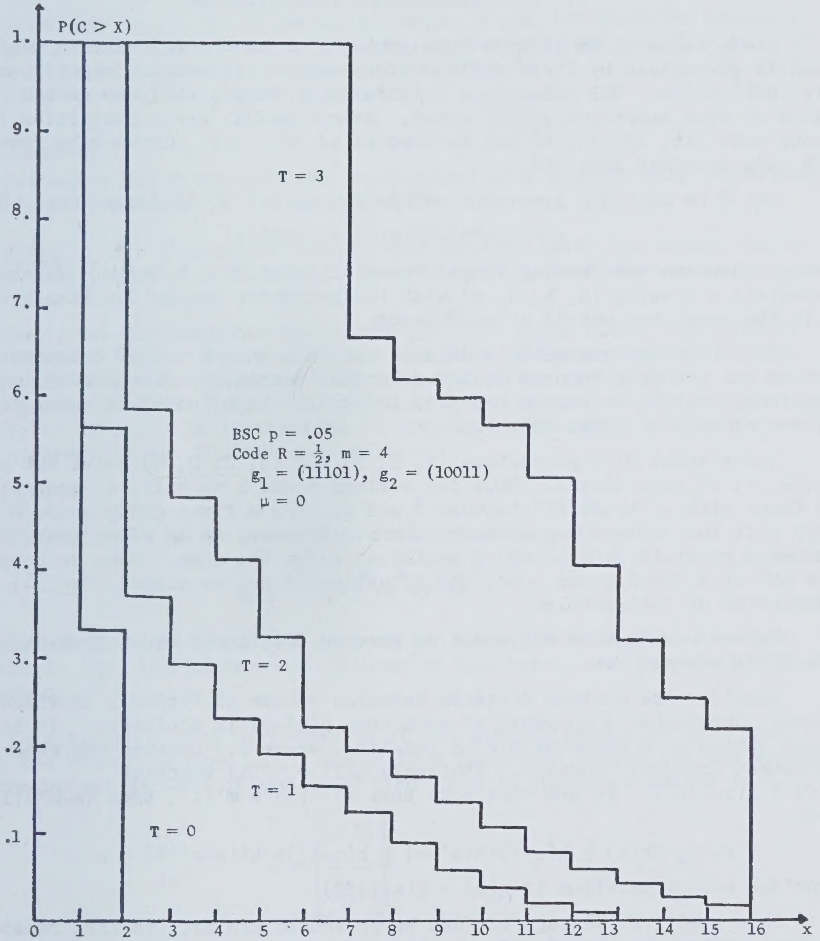


Figure 5. Experimental distributions of the number of computations per decoded bit for the Adaptive Reduced-State Viterbi decoder.

BINARY MOST SIGNIFICANT BIT CODES

LARRY A. DUNNING and W. E. ROBBINS
North Carolina State University
Raleigh, North Carolina

ABSTRACT

Systematic most significant bit (group) codes are characterized. Several constructions for goup codes are generalized to utilize and/or produce MSB codes. Some low rate group codes with larger minimum distances than those previously reported are produced.

I. INTRODUCTION AND PRELIMINARIES

Group codes which provide more protection for one information digit than is guaranteed by their minimum distance are called most significant bit (MSB) codes. MSB codes were introduced by Kilgus [14] who gave a table of some short cyclic MSB codes. Every unequal error protection (UEP) group code [15, 13, 14, 4] can be used as an MSB code. Conversely, every MSB code provides some UEP.

Let G be an k x n generator matrix for an (n, k, d) group code. Define

$$s = \min\{h(mG):m(1) = 1\varepsilon GF(2)\}$$

where h denotes the Hamming weight function and m is a k-tuple. We shall associate a 4-tuple (n, k, d, s) with the generator matrix G. When $s > d$, G is the generator matrix of an MSB-code.

Note that the parameter s depends not only on the set of codewords, but on the encoding induced by the generator matrix, G. The reader is cautioned that it is common practice to use the term "code" to refer to either or both of these objects.

Group codes with parameters (5, 2, 3) and (5, 2, 2, 4) exist but no (5, 2, 3, 4) code exists. Thus for a fixed n and k we will, in general, be faced with a "tradeoff" between d and s. For a fixed group code, s may vary with the choice of generator matrix. However, it is clear that one may choose a generator matrix which maximizes s for the code. When we consider the UEP of a fixed group code, the situation is not as clear. See [4] for a resolution of this problem.

We now verify that MSB codes do provide additional error protection for the first message bit.

Consider the minimum distance decoding scheme of Figure 1 in which the decoder determines a codeword c' such that h(c'-r) is minimized. It is well known that m' = m whenever $h(e) \le \lfloor (d-1/2 \rfloor$, where $\lfloor . \rfloor$ denotes the floor (greatest integer) function. Similarly m(1) = m'(1) whenever $h(e) \le \lfloor (s-1)/2 \rfloor$. To see this note that if $m(1) \ne m'(1)$, then $(m-m')(1) = 1$ and

$$s-1 \ge 2h(e) \ge h(c-r)+h(c'-r) \ge h(c-c')= h((m-m')G) \ge s$$

provides a contradiction if $h(e) < \lfloor (s-1)/2 \rfloor$.

Thus the first message bit can be protected against $\lfloor (s-1/2 \rfloor$ random errors. Other results of this type are possible. For example, G may be used to protect the first message bit against s - 1 erasures.

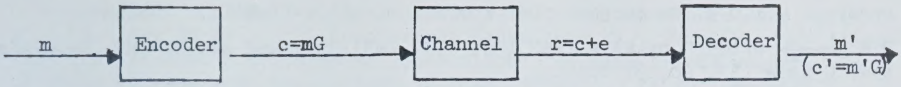


Figure 1.

II. THE ROLE OF NONSYSTEMATIC CODES

Gore and Kilgus [5] gave the first example of a nonsystematic MSB group code. They showed that the (15, 9, 4) code of Table 1 has s = 5. This code is majority logic decodable. However, a systematic (15, 9, 4, 5) code may be obtained from a (14, 9, 4) code by means of the following construction.

Given any (n, k, d, s) generator matrix G and any positive integer, p, we may construct a generator matrix with parameters (n+p, k, d, s + p) by adding to G, as columns, p k-tuples of the form (1, 0, ..., 0). We shall call this construction "Construction T."

Proposition 1:

Any systematic MSB group code can be obtained from a group code of shorter block length by means of "Construction T".

Proof: Such a generator matrix must have at least one column vector of the form (1, 0, ..., 0) which may be dropped to form the requisite code of shorter block length. Q.E.D.

Thus, for the construction of MSB (or UEP) group codes, the authors favor methods such as those of Kilgus and Gore [14] which result in nonsystematic codes over those of Masnick and Wolf [15] which result in systematic codes. Gore and Kilgus [5] obtained a specialized result of this type for cyclic codes. The limitations of systematic MSB codes are made even more apparent by the following example.

Example 1:

The (35, 28) cyclic group code with generator polynomial

$$g(X) = 1+X^{3}+X^{4}+X^{6}+X^{7}$$

has minimum distance 4 and can protect one message bit against 2 errors. No systematic (35, 28) binary code, linear or nonlinear, can provide equivalent protection.

Verification: For verification of the minimum weight of this code we refer the reader to Chen [3]. Let G' be a generator matrix for the subspace of even weight vectors of the code. Our 28 x 35 generator matrix is G.

$$G = \begin{bmatrix} g(X) \\ \text{----} \\ G \end{bmatrix}$$

Whenever m(1) = 1, mG is easily seen to have odd weight. It follows that s ≥ 5. And, h(g(X)) = 5 so, in fact, s = 5.

Now suppose $m \longrightarrow mN \in GF(2)^{35}$ where $m \in M = GF(2)^{28}$ is a systematic code with minimum distance 4 which protects its first information bit against two errors. w.ℓ.o.g. we suppose that for all m, m(1) = (mN)(1). Define $E:M \longrightarrow GF(2)^{33}$ by mE = ((mN)(3), ..., (mN)(35)) and set C = ME. We claim that C is a (33, 28, 3) code.

Let $m,m' \in M$. If m(1) = m'(1), then (mN)(1) = (m'N)(1) and mN and m'N differ in at least 4 positions. (mN)(2) = (m'N)(2) could account for at most one of these, so mE and m'E differ in at least three positions. On the other hand, suppose m(1) = m'(1); mN and m'N must differ in at least 5 positions lest N not protect its first message bit against 2 errors. $(mN)(1) \neq (m'N)(1)$ accounts for one position. $(mN)(2) \neq (m'N)(2)$ could account for at most one more. Thus mE and m'E differ in at least three positions in this case too. We have also shown that E is 1 - 1.

Thus C is a (33, 28, 3) code and has 2^{28} = 268,435,456 codewords. According to Johnson [12] a binary code with block length 33 and minimum distance 3 has at most 2.526×10^8 codewords. Our contradiction is apparent.

III. CONSTRUCTIONS

We now examine several constructions which utilize and/or produce MSB codes. We shall seek MSB codes which are also "good" group codes and cannot be obtained by means of Construction T. For example, the (35, 28, 4, 5)code of the last section has the property that no group code with k=28 and d=4 has smaller block length [10].

We define minimum odd weight, d(odd), of a group code to be the minimum weight of the odd-weight codewords of the code. If the code fails to have odd-weight codewords, set d(odd) = $+\infty$. The minimum even weight of a group code, d(even) is defined similarly, except that the zero codeword is not considered when the minimum is taken.

Let d(odd) denote the minimum odd weight of a particular (n, k, d)group code. Most known MSB codes can be based on the following observation. Any generator matrix for this code with first row of odd weight and all others of even weight has parameters (n, k, d, d(odd)).

Several authors [13, 9, 8, 7] have developed methods which can be used to determine the minimum odd weight of some cyclic codes of composite length. Selected cyclic MSB codes, obtained from these sources, are given in Table 1. The following theorem has been used to establish (lower bounds on) the minimum odd weight of such codes. The form given here is a generalization of Theorem 2 of [7].

Theorem 1:

Let two group codes C1 and C2 be given. Denote the minimum distance, minimum odd weight, and minimum even weight of Ci by d(i), d(iodd), and d(ieven) respectively, i = 1, 2. Then the minimum odd and minimum even weights of the product code formed using these codes are given by:

$$d(odd) = d(1odd)\cdot d(2odd);$$
$$d(even) = \min(d(1even)d(2),\ d(1)d(2even)).$$

Proof: We show only that $d(odd) \geq d(1odd)\cdot d(2odd)$; $d(odd) \leq d(1odd)\cdot d(2odd)$ is easy. The expression for d(even) will not be of interest to us but is included for completeness. Its proof is similar.

Denote the product code by C. Let $c \in C$ be arbitrary with h(c) odd. Let n2 be the block length of C2 and set $c' = \Sigma\{c(i,*): 1 \leq i \leq n2\}$ where c(i,*) denotes the i^{th} row of c. c'(j)=1 iff c(*,j) has odd weight where c(*,j) denotes the j^{th} column of c. The weight of c' must then be odd lest c have an even number of columns of odd weight and hence have even weight. c' ε C1 since it is a sum of codewords taken from C1. Thus, $h(c') \geq d(1odd)$. This implies that c contains at least d(1odd) columns of odd weight. Each of these columns must have weight at least d(2odd). It follows that $h(c) \geq d(1odd)d(2odd)$ which completes the proof. Q.E.D.

Table 1: Selected Cyclic MSB Codes

n	k	d	d(odd)	gen. poly. roots	sources
15	9	4	5	3,5	[14,2]
21	13	4	7	3,7,9	[14,2]
33	13	10	11	1,3	[13,3]
35	28	4	5	5,7	[3],Ex.1
39	13	12	15	1,3,13	[13,3]
39	15	10	13	1,3	[13,3]
51	17	16	19	1,3,5,9,17	[2]
51	19	14	17	1,3,5,9	[13,3]
51	27	8	17	1,3,9	[7]
57	21	14	19	1,3	[13,3]
87	31	22	29	1,3	[8]
89	23	28	29	1,5,9,11,19,33	[11]
91	13	36	39	3,7,13,17,19,39	[2]
111	39	22	37	1,3	[8]

The following theorem is a generalization of a construction due to Plotkin, Sloane, and Whitehead [16, 18]. It may be used to construct some MSB codes which are also "optimal" group codes; e.g., a (10, 5, 4, 5) group code may be constructed. Selected MSB group codes obtained with this construction are given in Table 2.

Theorem 2:

If there exist $(n,k, \geq d, \geq s)$ and $(n,k', \geq d')$ group codes with $d' \geq (1/2)d$, then there exists a $(2n, k + k', \geq d, \geq s)$ group code.

Proof: Let A be an $(n, k, \geq d, \geq s)$ generator matrix, and let B be an $(n, k', \geq d')$ generator matrix with $d' \geq (1/2)d$. It is not difficult to verify that

$$G = \left[\begin{array}{c|c} A & 0 \\ \hline B & B \end{array}\right]$$

is a $(2n, k + k', > d, > s)$ generator matrix. The verification is a simple generalization of that found in [18]. Q.E.D.

The construction X of Sloane, Reddy and Chen [17] can utilize MSB codes. This is illustrated by the following example.

Example 2:

We shall construct a (36, 13, 12) group code. Let G' be a generator matrix for the (33, 10, 12) cyclic code with roots $1, \alpha, \alpha^3, \alpha^{11}$. Let v1, v2 be row vectors such that the rows of G' together with v1 and v2 generate the (33, 12, 10) cyclic code with roots $1, \alpha, \alpha^3$. Now let $\bar{1}$ denote the row vector of 33 ones. The rows of G' together with v1, v2 and $\bar{1}$ generate the (33, 13, 10) cyclic code with roots α and α^3. The minimum odd weight of this last code is 11. The information at hand is sufficient to prove that

$$G = \left[\begin{array}{c|ccc} \bar{1} & 0 & 0 & 1 \\ v1 & 1 & 0 & 1 \\ v2 & 0 & 1 & 1 \\ \hline G' & & 0 & \end{array}\right]$$

is the generator matrix of a (36, 13, 12) group code. Note that the weight of columns 1 - 33 of a codeword is odd iff m(1)=1.

Table 2: Selected MSB Codes

n	k	d	s	source
$10=2\cdot5$	$5=1+4$	$4=2\cdot2$	5	Th. 2
$14=2\cdot7$	$5=1+4$	$6=2\cdot3$	7	Th. 2
$18=2\cdot9$	$12=4+8$	$4=2\cdot2$	5	Th. 2
$22=2\cdot11$	$5=1+4$	$10=2\cdot5$	11	Th. 2
$30=2\cdot15$	$6=1+5$	$20=2\cdot10$	21	Th. 2
$34=2\cdot17$	$27=11+16$	$4=2\cdot2$	5	Th. 2
$36=2\cdot18$	$7=1+6$	$16=2\cdot8$	18	Th. 2
$38=2\cdot19$	$8=1+7$	$16=2\cdot8$	19	Th. 2
$40=2\cdot20$	$9=1+8$	$16=2\cdot8$	20	Th. 2
$42=2\cdot21$	$6=1+5$	$20=2\cdot10$	21	Th. 2
$42=(31+1)+10$	$11=6+5$	$16=12+4$	$17=12+5$	const. X
$54=51+3$	$19=17+(1+1)$	$16=14+2$	17	const. X
$62=2\cdot31$	$7=1+6$	$30=2\cdot15$	31	Th. 2
$66=2\cdot33$	$58=26+32$	$4=2\cdot2$	5	Th. 2
$100=2\cdot50$	$8=1+7$	$48=2\cdot24$	50	Th. 2
$102=2\cdot51$	$9=1+8$	$48=2\cdot24$	51	Th. 2
112	8	52	≥56	IM_3 of [1]

The following theorem was suggested by Theorem 1 of Griesmer [6]. The notation $\lceil . \rceil$ denotes the "ceiling" function, $\lceil x \rceil = - \lfloor -x \rfloor$.

Theorem 3:

If there exists an $(n, k, \geq d, \geq s)$ group code, then there exists an $(n - s, k - 1, \geq \lceil d/2 \rceil)$ group code.

Proof: Let G be a generator matrix for a $(n, k, \geq d, \geq s)$ group code.

Let C be the vector space spanned by the last k-1 rows of G. Assume w.ℓ.o.g. that the first row G(1,*) of G has minimal weight in G(1,*)+C. G(1,*) has at least s ones. Assume w.ℓ.o.g. that the first s positions of G(1,*) are ones. Let C' denote the code generated by the generator matrix G' obtained by removing the first row and the first s columns from G. We claim that C' is a $(n-s, k-1, \geq \lceil d/2 \rceil)$ group code. This follows since no $c \in C$ can have more than 1/2 h(c) ones in the first s positions lest $h(c + G(1,*)) < h(G(1,*))$. Thus at least $\lceil (1/2)h(c) \rceil \geq \lceil d/2 \rceil$ ones are in the remaining positions. Q.E.D.

We could delete n columns corresponding to a codeword of weight d in a $(n, k, d, \geq s)$ group code and by similar argument obtain an

$(n-d, k-1, \geq \lceil d/2 \rceil, \geq s - \lfloor d/2 \rfloor)$ group code. However, these parameters can be obtained using the code guaranteed by the last theorem and applying "Construction T". One adds s - d additional columns and notes that $(s-d)+\lceil d/2 \rceil = s- \lfloor d/2 \rfloor$.

The preceding theorem may be used to obtain upper bounds for MSB group codes. For example, no $(\leq 9, 4, 3, 5)$ group code exists lest a $(\leq 4, 4, 2)$ group code exist. This tightens a lower bound given by Masnick [15].

Given (n, k, d) and(n', k, d') group codes it is well known that we may concatenate their generator matrices to form the generator matrix of a $(n + n', k, \geq d + d')$ group code. This result is generalized to utilize and/or produce MSB group codes in the following theorem. The constructions obtained yield several low rate group codes with minimum distances better than those reported by Helgert and Stinaff [10]. A listing is given in Table 4.

Theorem 4:

Suppose there exist $(ni, k, \geq di, \geq si)$ group codes, i = 1, 2. Then,

(i) There exists an $(n1+n2, k, \geq d1+d2, \geq s1+s2)$group code.

(ii) If s1 > d1, then there exists an $(n1+n2-1, k, \geq d1+d2, \geq s1+d2-1)$ group code.

(iii) If s1 > d1 and s2 > d2, then there exists an $(n1+n2-2, k, \geq d1+d2, \geq s1+d2-1)$ group code.

Proof: We give only the proof of the most difficult part, (iii), which utilizes all the techniques necessary for (i) and (ii).

Let Gi be a generator for a matrix for a $(ni, k, \geq di, \geq si)$ group code, i = 1,2. We assume $k \geq 2$. It is not difficult to show that we may assume (w.ℓ.o.g.) that Gi has the form shown.

$$Gi = \left[\begin{array}{c|c} 0 & \\ 0 & \\ \vdots & Hi \\ 0 & \\ 1 & \end{array}\right] \qquad G = \left[\begin{array}{c|c} H1 & H2 \end{array}\right]$$

for some k x (n-1) matrix Hi, i=1, 2. Let K2 be a kx(n-1) matrix defined by K2(i,j)=H2(k-i,j). That is, K2 is H2 turned upside down. One may verify that the generator matrix G has parameters $(n1+n2-2,k, \geq d1+d2, \geq s1+d2-1)$ by verifying the entries in the following table. Note that $h(mG)=h(mH1)+h(mK2)$ for all $m \in GF(2)^k$. H1 and/or K2 will not necessarily have maximal row rank, but G will.

Table 3

m(1)	m(k)	h(mH1)	H(mKs)	h(mG)
0	0	≥d1	≥d2	≥d1+d2
0	1	≥d1-1	≥s2≥d2+1	≥d1+d2
1	0	≥s1	≥d2-1	≥s1+d2-1
1	1	≥s1-1	≥s2-1≥d2	≥s1+d2-1

Q.E.D.

Table 4: Improved Low Rate Group Codes

N	k	d	s	comments
73=42+32-1	6	36	-	
77=30+48-1	6	38	-	
88=67+27-1	7	42	-	known [1]
107=100+8-1	8	49	50	
110=102+9-1	9	49	51	
111=100+12-1	8	51	52	known [1]
114=102+13-1	9	51	53	
115=100+16-1	8	53	54	

REFERENCES

[1] W. O. Alltop, "Binary codes with improved minimum weights," IEEE Trans. Inform. Theory, vol. IT-22, pp. 241-243, March 1976.

[2] E. R. Berlekamp, Algebraic Coding Theory, New York: McGraw-Hill, 1968, pp. 432-433.

[3] C. L. Chen, "Computer results on the minimum distance of some binary cyclic codes," IEEE Trans. Inform. Theory, vol. IT-16, pp. 359-360.

[4] L. A. Dunning,"Encodings, UEP, and sparse generator matrices for linear block codes," Ph.D. dissertation, Dept. of Mathematics, N. C. State Univ., Raleigh, May 1976.

[5] W. C. Gore and C. C. Kilgus, "Cyclic codes with unequal error protection," IEEE Trans. Inform. Theory, vol. IT-17, pp. 215-215, March 1971.

[6] J. H. Greismer, "A bound for error-correcting codes," IBM J. Res. Development, vol. 4, pp. 532-542, Nov. 1960.

[7] C. R. P. Hartmann and T. Y. Hwang, "Some Results on the Weight structure of cyclic codes of composite length," IEEE Trans. Inform. Theory, vol. IT-22, pp. 340-348, May 1976.

[8] C. R. P. Hartmann and K. K. Tzeng, "On some classes of cyclic codes of composite length," IEEE Trans. Inform. Theory, vol. IT-19, pp. 820-823, Nov. 1973.

[9] C. R. P. Hartmann, K. K. Tzeng, and R. T. Chien, "Some results on the minimum distance structure of cyclic codes," IEEE Trans. Inform. Theory, vol. IT-18, pp. 402-409, May 1972.

[10] H. J. Helgert and R. D. Stinaff, "Minimum-distance bounds for binary linear codes," IEEE Trans. Inform. Theory, vol. IT-19, pp. 344-356, May 1973.

[11] H. J. Gelgert and R. D. Stinaff, "Shortened BCH codes," IEEE Trans. Inform. Theory, vol. IT-19, pp. 818-820, Nov. 1973.

[12] S. M. Johnson, "On upperbounds for unrestricted binary error-correcting codes," IEEE Trans. Inform. Theory, vol. IT-17, pp. 466-478, July 1971.

[13] T. Kasami, "Some lower bounds on the minimum weight of cyclic codes of composite length," IEEE Trans. Inform. Theory, vol. IT-14, pp. 814-818, Nov. 1968.

[14] C. C. Kilgus and W. C. Gore, "A class of cyclic unequal-error-protection codes," IEEE Trans. Inform. Theory, vol. IT-18, pp. 687-690, Sept. 1972.

[15] B. Masnick and J. Wolf, "On linear unequal error protection codes," IEEE Trans. Inform. Theory, vol. IT-13, pp. 600-607, Oct. 1967.

[16] M. Plotkin, "Binary codes with specified minimum distance," IEEE Trans. Inform. Theory, vol. IT-16, pp. 445-450, Sept. 1960.

[17] N. J. A. Sloane, S. M. Reddy and C. L. Chen, "New binary codes," IEEE Trans. Inform. Theory, vol. IT-18, pp. 503-510, July 1972.

[18] N. J. A. Sloane and D. S. Whitehead, "New family of single-error correcting codes," IEEE Trans. Inform. Theory, vol. IT-16, pp. 717-719, Nov. 1970.

ON UNIQUELY DECIPHERABLE CODES WITH GIVEN COMPOSITIONS[1]

S. C. NTAFOS
Department of Electrical Engineering
Northwestern University
Evanston, Illinois

S. L. HAKIMI
Departments of Electrical Engineering & Computer Sciences
Northwestern University
Evanston, Illinois

ABSTRACT

It has been conjectured that a uniquely decipherable (U.D.) code can be replaced by a prefix code with the same codeword compositions. The conjecture is proved for the two length case, that is for U.D. codes with codeword lengths from the set $\{\ell_1,\ell_2\}$. This result is then extended to a more general class of U.D. codes and an equivalent conjecture is proposed for the general case.

1. Introduction

The basic elements of a discrete communication system are its source, encoder, channel, decoder and destination. The source may be represented as a random variable, X, taking on values from a source alphabet $\{x_1,x_2,\ldots,x_M\}$ with propabilities $p_1,p_2,\ldots,p_M$ respectively. A message is a sequence of source characters. To facilitate transmission, the encoder associates with every source character, x_i, a finite sequence of code characters from the code alphabet $\{a_1,a_2,a_3,\ldots,a_D\}$. Such a sequence of code characters is called a codeword. A code is the collection of all codewords. The encoded message is then transmitted over the channel which we assume to be noiseless. At the receiving end, the decoder attempts to reproduce the original message by assigning a set of source characters to the coded message.

To avoid ambiguity, every finite sequence of code characters must correspond to no more than one message. A code that conforms with this requirement is said to be a uniquely decipherable (U.D.) code. Furthermore, if no codeword is a prefix to some other codeword, the code is said to be a prefix (or instantaneous) code. A codeword w_j, is a prefix of codeword w_i, if $w_i=w_jb_j$, where w_j, b_j are in juxtaposition and b_j is some sequence of code characters.

The composition of a codeword w_i, is written as $(d_1(i),d_2(i),\ldots,d_D(i))$ where $d_j(i)$ is the number of times code character a_j appears in codeword w_i. Given a set of costs $(c_1,c_2,\ldots,c_D)$ associated with the respective code characters $\{a_1,a_2,\ldots,a_D\}$ the code cost is given by:

---

[1] This work is supported by the U.S. Air Force Office of Scientific Research, Systems Command, Grant AF-AFOSR-76-3017.

$$\text{Code cost} = \sum_{i=1}^{M} \cdot \sum_{j=1}^{D} p_i d_j(i) c_j$$

The problem of minimizing the code cost has been treated for prefix codes. Karp [2] gave an integer programming formulation and, if all the costs are equal, the problem is solved by the well-known Huffman algorithm [3]. However, no results exist for uniquely decipherable codes. It is known that a U.D. code can be replaced by a prefix code with the same codeword lengths.

Theorem 1: (Kraft [4]). A prefix code with codeword lengths $n_1, n_2, \ldots, n_M$ exists if, and only if, the following condition holds:

$$\sum_{i=1}^{M} D^{-n_i} \le 1$$

where D is the number of characters in the code alphabet.

McMillan [5] showed that the same inequality must hold for uniquely decipherable codes. Thus, every U.D. code can be replaced by a prefix code with the same codeword lengths.

If the costs $(c_1, c_2, \ldots, c_D)$ are not all equal, the composition of the codewords becomes important. The following theorem, independently discovered by Block [6] and Carter and Gill [7], establishes necessary and sufficient conditions for the existence of a prefix code with a given set of compositions.

Theorem 2: Given a set of compositions of the form $C = \{(m_1, m_2, \ldots, m_D)\}$, there exists a prefix code with these exact codeword compositions if, and only if, the following inequality is satisfied for each $(m_1, m_2, \ldots, m_D) \in C$.

$$\prod_{k=1}^{D-1} \binom{\sum_{i=k}^{D} m_i}{m_k} \ge \sum_{l=1}^{\sum_{i=1}^{D} m_i} \quad \sum_{\substack{n_1+n_2+\ldots=l \\ n_i \le m_i}} \left\{ \prod_{k=1}^{D-1} \binom{\sum_{i=k}^{D} (m_i - n_i)}{m_k - n_k} \right\} [n_1, n_2, \ldots, n_D]$$

where $[n_1, n_2, \ldots, n_D]$ is the number of codewords with composition $(n_1, n_2, \ldots, n_D)$.

A summary of the proof is described as follows:

The necessity of the condition is a direct result of the prefix property. Let a word be a sequence of code characters. Then the product term on the right side represents the number of words of composition $(m_1, m_2, \ldots, m_D)$ that have as prefix a word with composition $(n_1, n_2, \ldots, n_D)$. The product term on the left side is the number of all words of composition $(m_1, m_2, \ldots, m_D)$. The inequality states that the number of words of composition $(m_1, m_2, \ldots, m_D)$ must be greater than or equal to the number of words of composition $(m_1, m_2, \ldots, m_D)$ that are eliminated by the prefix property, plus the number $[m_1, m_2, \ldots m_D]$.

The sufficiency of the condition can be shown by constructing a prefix code with the given compositions. The construction proceeds from the shorter to the longer codewords. At each step, if we require $[m_1,m_2,\ldots,m_D]$ words of composition $(m_1,m_2,\ldots,m_D)$, it follows from the composition inequalities that there are at least $[m_1,m_2,\ldots,m_D]$ words of composition $(m_1,m_2,\ldots,m_D)$ and such that they do not have a prefix in the set of words chosen so far. The code, thus constructed, will be a prefix code with the required codeword compositions.

Example 1. Given the compositions (1,1),(1,1),(2,0),(1,2), is there a binary prefix code whose codewords have the given compositions?

The composition inequalities are:

$$\text{for } (1,1); \quad \binom{1+1}{1} \geq \binom{(1-1)+(1-1)}{1-1}[1,1] = \binom{0}{0}2$$

$$\text{or} \quad 2 \geq 2$$

$$\text{for } (2,0); \quad \binom{2+0}{2} \geq \binom{(2-2)+(0-0)}{2-2}[2,0] = \binom{0}{0}1$$

$$\text{or} \quad 1 \geq 1$$

$$\text{for } (1,2); \quad \binom{1+2}{1} \geq \binom{(1-1)+(2-1)}{1-1}[1,1] + \binom{(1-1)+(2-2)}{1-1}[1,2]$$

$$\text{or} \quad 3 \geq 3$$

All the composition inequalities are satisfied. The only prefix code with the given codeword compositions is: $W = \{01,10,00,110\}$. However, if one more composition (1,2) is added to the above set, the last inequality becomes:

$$\binom{1+2}{1} \geq \binom{(1-1)+(2-1)}{1-1}[1,1] + \binom{(1-1)+(2-2)}{1-1}[1,2]$$

$$\text{or} \quad 3 \geq 4$$

Thus, there is no prefix code with codeword compositions $\{(1,1),(1,1),(2,0),(1,2),(1,2)\}$. This is easily confirmed by constructing a binary tree.

It has been conjectured [6,7] that theorem 2 holds for U.D. codes. If this is true, the problem of constructing minimum cost uniquely decipherable codes reduces to that of constructing minimum cost prefix codes.

2. The Two Length Case

In this section we show that every uniquely decipherable code with codeword lengths from the set $\{\ell_1,\ell_2\}$ can be replaced by a prefix code with the same codeword compositions.

Let $W^2 = \{w_1,w_2,\ldots,w_M\}$ be a non-prefix uniquely decipherable code such that $W^2 = L_1 \cup L_2$ where:

$$L_1 = \{w_i \mid \lambda(w_i) = \ell_1\}$$

$$L_2 = \{w_j \mid \lambda(w_j) = \ell_2 > \ell_1\}$$

and $\lambda(w)$ is the length of codeword w.

Lemma 1. Let W be a non-prefix U.D. code, $w_i, w_j \in W$ and $w_j = w_i b_i$. Then $b_i w_i \notin W$.

Proof. Suppose that $b_i w_i = w_k \in W$. Consider the message $w_j w_i$. This message may be written as:

$$w_j w_i = w_i b_i w_i = w_i w_k$$

Hence, there would be two interpretations for the same message (i.e. $w_j . w_i$ or $w_i . w_k$) and the code is not a U.D. code. This contradicts our assumption that W is a U.D. code. Thus, $b_i w_i \notin W$.

Consider now the code W^2. For each codeword $w_j \in W^2$ that has a prefix in W^2, we construct a codeword chain, A_{w_j}. The first word in the chain is $w_j = w_i b_i$; the next word is $b_i w_i$. If $b_i w_i$ does not have a prefix in W^2 the chain terminates with $b_i w_i$. Otherwise, we can write $b_i w_i = w_k b_k$ ($w_k \in W^2$) and the next word in the chain is $b_k w_k$. This process is repeated until a word that has no prefix in W^2 is reached. It will be shown later that such a chain must terminate. We represent the chain associated with $w_j = w_i b_i$ by:

$$A_{w_j} = w_j = w_i b_i \rightarrow b_i w_i = w_k b_k \rightarrow b_k w_k = \ldots. = w_x b_x \rightarrow b_x w_x = \ldots.$$

Example 2. Let $W^2 = \{000,001,010,100,00101\}$. The codeword 00101 has the prefix 001 and the chain A_{00101} is given by:

$$A_{00101} = 00101 \rightarrow 01001 \rightarrow 01010 \rightarrow 10010 \rightarrow 10100$$

Clearly, all words in a chain A_{w_j}, have the same composition as they are cyclic permutations of w_j.

Theorem 3. Let $w_j \in W^2$, $w_j = w_i b_i$ ($w_i \in W^2$) and let A_{w_j} be the chain associated with w_j. Then, the first word of A_{w_j} is the only member of the chain that is a codeword of W^2.

Proof. Let A_{w_j} be:

$$A_{w_j} = w_j = w_i b_i \rightarrow b_i w_i = w_k b_k \rightarrow b_k w_k = w_1 b_1 \rightarrow b_1 w_1 = \ldots. = w_x b_x \rightarrow b_x w_x = \ldots$$

Suppose that b_xw_x is a member of the chain and also a codeword of W^2. Consider the message $B = w_jw_iw_kw_1\dots w_{x-1}w_x$. Since $w_j = w_ib_i$, we can write:

$$B = w_ib_iw_iw_kw_1\dots w_{x-1}w_x$$

$$\begin{array}{lll}
\text{but } b_iw_i = w_kb_k & \Rightarrow & B = w_iw_kb_kw_kw_1\dots w_{x-1}w_x \\
\text{but } b_kw_k = w_1b_1 & \Rightarrow & B = w_iw_kw_1b_1w_1\dots w_{x-1}w_x \\
\vdots & & \vdots \\
\vdots & \Rightarrow & B = w_iw_kw_1\dots b_{x-1}w_{x-1}w_x \\
\text{but } b_{x-1}w_{x-1} = w_xb_x & \Rightarrow & B = w_iw_kw_1\dots w_{x-1}w_xb_xw_x
\end{array}$$

Now, if $b_xw_x \in W^2$, there would be two interpretations for message B. That is $B = w_j\cdot w_i\cdot w_k\dots w_x = w_i\cdot w_k\cdot w_1\dots w_x\cdot b_xw_x$. This contradicts our assumption that W^2 is a U.D. code. Therefore, $b_xw_x \notin W^2$ for any x.

We noted before that a chain must terminate. If a chain A_{w_j} does not terminate, because of the finite number of words of composition equal to that of w_j, the chain must cycle. In other words, at some point a word will be obtained that has already appeared in the chain. Then part of, or the whole chain, will be repeated indefinitely.

Theorem 4. Let $w_i, w_j \in W^2$ and $w_j = w_ib_i$. Let A_{w_j} be the chain associated with w_j. Then A_{w_j} contains no cycles.

Proof. First suppose that the chain cycles to a word other than w_j. That is:

$$A_{w_j} = w_j = w_ib_i \to b_iw_i = \dots = w_nb_n \to b_nw_n = \dots = w_pb_p \to b_pw_p = b_nw_n$$

Since the codewords w_n and w_p are of equal length (ℓ_1), it follows that $w_n = w_p$ and $b_n = b_p$. But then from the chain we have that:

$$b_{n-1}w_{n-1} = w_nb_n \text{ and } b_{p-1}w_{p-1} = w_pb_p$$

Since $w_nb_n = w_pb_p$ it follows that $b_{n-1}w_{n-1} = b_{p-1}w_{p-1}$. Then, by the same argument, $w_{n-1} = w_{p-1}$ and $b_{n-1} = b_{p-1}$. This process is repeated until we reach w_j. Then we have the following situation:

$$A_{w_j} = w_j = w_ib_i \to b_iw_i = \dots\dots = w_1b_1 \to b_1w_1 = w_j$$

Theorem 3 states that the first word of A_{w_j} (i.e. w_j) is the only word that belongs to W^2. But if $b_1w_1 \notin W^2 \Rightarrow w_j \notin W^2$, a contradiction. The ambiguous message for this case is:

$$B = w_j \cdot w_i \ldots\ldots\ldots w_1 = w_i \cdot w_k \ldots\ldots\ldots w_j .$$

We now describe a method for replacing a uniquely decipherable code W^2 with a prefix code with the same codeword compositions as W^2. Given the U.D. code W^2, define the set S as follows:

$$S = \{w_j \in W^2 \mid w_j = w_i b_i \text{ for some } w_i \in W^2\}$$

Then, for each member of S, we construct a chain and form the set S' where:

$$S' = \{b_x w_x \mid w_x \in W^2 \text{ and } b_x w_x \text{ is the last word in a chain } A_{w_j},\ w_j \in S\}$$

Lemma 2. If $|C|$ is the cardinality of set C, we have $|S| = |S'|$.

Proof. This follows from the fact that no two chains can have any words in common. Suppose that the chains A_{w_k}, A_{w_j} have common words. Then,

$$A_{w_j} = w_j = w_1 b_1 \rightarrow b_1 w_1 = \ldots\ldots = w_i b_i$$

$$b'_f w'_f = b_i w_i = w_{i+1} b_{i+1} \rightarrow \ldots\ldots$$

$$A_{w_k} = w_k = w'_1 b'_1 \rightarrow b'_1 w'_1 = \ldots\ldots = w'_f b'_f$$

By the argument used in the proof of Theorem 4, we have: $w_i = w'_f$, $b_i = b'_f$ which implies that $w_{i-1} = w'_{f-1}$, $b_{i-1} = b'_{i-1}$ and so on. The two chains cannot have equal lengths because that would require that $w_i = w_k$. Thus, the shorter of the two chains, say A_{w_k}, is a part of the longer one, A_{w_j}. But then the codeword $w_k \in W^2$ appears in A_{w_j}, while, by Theorem 3, w_j is the only member of A_{w_j} that belongs to W^2. Thus $w_k \notin W^2$, a contradiction.

The code $W = W^2 - S + S'$ is then a prefix code with the exact same codeword compositions as W^2.

Example 3. Consider the binary U.D. code W given by:

$$W^2 = \{000,001,010,100,00110,10011,01011,01111,11011,00101\}$$

The set S is: $S = \{00101,00110,10011,01011\}$

We now construct the chains associated with each member of S.

$$A_{00101} = 00101 \rightarrow 01001 \rightarrow 01010 \rightarrow 10010 \rightarrow 10100$$

$$A_{00110} = 00110 \rightarrow 10001 \rightarrow 01100$$

$$A_{10011} = 10011 \rightarrow 11100$$

$$A_{01011} = 01011 \rightarrow 11010$$

Then the set S' is: $S' = \{10100,01100,11100,11010\}$ and the prefix code W' is:

$$W' = \{000,001,010,100,10100,01100,11100,11010,01111,11011\}$$

3. An Extension of the Two Length Case

The method used for the two length case can be applied to a more general class of U.D. codes. The basic requirements are that the concept of a codeword chain is applicable and that Lemma 2 holds. Let W_p be a non-prefix U.D. code such that:

$$W_p = \{w_1,w_2,w_3,\ldots,w_M\} = P_1 \cup P_2$$

$$P_1 = \{w_i \mid \lambda(w_i) = \ell, \quad \ell > \lambda(w_j) \ \forall\, w_j \in P_2\}$$

and P_2 is both a prefix and a suffix code. A code P is said to be a suffix code if no codeword $w_i \in P$ is a suffix to some other codeword $w_j \in P$, i.e. $w_j \neq b_i w_i$ for any $w_i, w_j \in P$.

Lemma 3. All suffix codes are uniquely decipherable.

Proof. Suppose that suffix code $P = \{w_1,w_2,\ldots,w_M\}$ is not uniquely decipherable. Then there is at least one ambiguous message. Let that message be: $B = w_1w_2w_3\ldots w_f = w_1'w_2'w_3'\ldots w_g'$. Then we can either write $w_f = b_g' w_g'$ or $w_g' = b_g w_f$ for some sequences of code characters b_f, b_g'. Either case contradicts our assumption that P is a suffix code.

Prefix and suffix codes are very similar in structure. However, suffix codes are impractical because decoding may have to wait until the whole message is received. As in the two length case, we can construct codeword chains A_{w_j} for all $w_j \in P_1$ that have a prefix in P_2. Also, Theorem 3 still holds.

Lemma 4. Let $A_{w_j}, A_{w_k}, \ldots A_{w_n}$ be the chains associated with the codewords $w_j, w_k, \ldots, w_n \in P_1$. Then

a) no two (or more) chains can have common words
b) all chains must terminate

Proof. a) From Theorem 3 it follows that no chain can be part of another chain. Also, since $w_j \neq w_k \neq \ldots \neq w_n$ and P_2 is a prefix code, at any step in the construction of the chains there is no more than one way for the chain to grow (i.e. a chain can not split up into two or more subchains). Suppose that chains A_{w_j}, A_{w_k} merge:

$$A_{w_j} = w_j = w_1b_1 \rightarrow b_1w_1 = \ldots = w_ib_i$$

$$b_iw_i = b_f'w_f' = w_gb_g \rightarrow b_gw_g = \ldots.$$

$$A_{w_k} = w_k = w_1'b_1' \rightarrow b_1'w_1' = \ldots = w_f'b_f'$$

From Theorem 3, it follows that $b_iw_i = b_f'w_f' \in P_1$. Then, $w_i, w_f' \in P_2$ and either w_i is a suffix of w_f', or w_f' is a suffix of w_i. This contradicts our assumption that P_2 is a suffix code.

b) If a chain does not terminate, it must cycle. Let A_{w_j} be:

$$A_{w_j} = w_j = w_1b_1 \rightarrow b_1w_1 = w_2b_2 \rightarrow b_2w_2 = \ldots = w_fb_f \rightarrow b_fw_f = \ldots = w_{g-1}b_{g-1} \rightarrow$$
$$\rightarrow b_{g-1}w_{g-1} = w_gb_g = w_fb_f$$

Since $w_g, w_f \in P_2$ and P_2 is a prefix code, we must have $w_g = w_f$. Also, since P_2 is a suffix code it follows that $w_{g-1} = w_{f-1}$ and so on until we reach w_j. Then, by Theorem 3, $w_j \in W_p$ a contradiction. Q.E.D.

Example 4. Consider the code W = {001,010,100,1010,00101}. We have:

$$A_{00101} = 00101 \rightarrow 01001 \rightarrow 01010 \rightarrow 10100 \rightarrow 01010 \rightarrow \ldots$$

The chain cycles. By Theorem 4b the code is not uniquely decipherable. One ambiguous message is:

$$B = 001 \cdot 010 \cdot 010 \cdot 100 \cdot 1010 = 00101 \cdot 00101 \cdot 001 \cdot 010$$

We have then shown that the U.D. code W_p can be replaced by a prefix code with the same codeword compositions.

Example 5. Consider the U.D. code

$$W_p = \{00,010,101,0110,1110,00111,10111,01001,11111\}$$

We have:
$$P_1 = \{00111,10111,01001,11111\}$$
$$P_2 = \{00,010,101,0110,1110\}$$

and
$$A_{00111} = 00111 \rightarrow 11100 \rightarrow 01110$$
$$A_{10111} = 10111 \rightarrow 11101 \rightarrow 11110$$
$$A_{01001} = 01001 \rightarrow 01010 \rightarrow 10010$$

The required prefix code is $W' = W_p - S + S'$ or

$$W' = \{00,010,101,0110,1110,01110,11110,10010,11111\}$$

4. The General Case

We start with the following conjecture.

Conjecture A. Every uniquely decipherable code must satisfy the composition inequalities established for prefix codes.

Suppose that conjecture A is false. Then there is a uniquely decipherable code such that at least one of the composition inequalities is not satisfied. Let $W^* = \{w_1, w_2, \ldots, w_M\}$ be such a code, and let $(m_1, m_2, \ldots, m_D)$ be a composition of minimum length,

$(\sum_{i=1}^{D} m_i)$, for which the corresponding inequality is not satisfied. Consider the code W defined by:

$$W = \{w_i \mid m_j(i) \le m_j,\ \exists\, j \ni m_j(i) < m_j,\ 1 \le j \le D\}$$

Code W is a U.D. code because it is a subcode of W^*. Furthermore, all the composition inequalities are satisfied for W. Therefore, there exists a prefix code, W', with the same codeword compositions as W. Let R be the set:

$$R = \{w_j \in W \mid k(w_j) = (m_1, m_2, \ldots, m_D)\}$$

where $k(w_j)$ is the composition of w_j. Then, the code WUR is a U.D. code but there is no prefix code with the same codeword compositions as WUR. This means that $|R|$ is greater than the number of words of composition $(m_1, m_2, \ldots, m_D)$ that can be added to W' without violating the prefix property. Equivalently, the U.D. code W eliminates less words of composition $(m_1, m_2, \ldots, m_D)$ than the prefix code W'. Thus, conjecture A is equivalent to:

Conjecture B. Given any non-prefix U.D. code $W = \{w_1, w_2, \ldots, w_M\}$ and a prefix code $W' = \{w'_1, w'_2, \ldots, w'_M\}$ such that $k(w_i) = k(w'_i) = (m_1(i), m_2(i), \ldots, m_D(i))$ for $1 \le i \le M$, let the composition $(m_1, m_2, \ldots, m_D)$ satisfy: $m_j \ge m_j(i)$ for all i,j and $\exists\, j \ni m_j > m_j(i)$, $i \le i \le M$, $1 \le j \le D$. Then code W eliminates at least as many words of composition $(m_1, m_2, \ldots, m_D)$ as code W'.

Code W' eliminates those words of composition $(m_1, m_2, \ldots m_D)$ that can be written as $w'_i\, b_j(i)$, $w'_i \in W'$. Consider the word pairs $(w_i b_j(i), b_j(i) w_i)$. From Lemma 1, it follows that at least one word from each pair is eliminated by W (if borh words are added to W, the augmented code is not a U.D. code). Since the number of pairs is equal to the number of words of composition $(m_1, m_2, \ldots, m_D)$ that are eliminated by W', it follows that conjecture B is true if the words $w_i b_j(i)$, $b_j(i) w_i$ are distinct for all i,j. Unfortunately, this is not the case. In certain cases, the matching that may occur, can be resolved by considering non-cyclic permutations. In general however, the question remains open.

5. Conclusion

It is proved that any uniquely decipherable code with codeword lengths from the set $\{\ell_1, \ell_2\}$ can be replaced by a prefix code with the same codeword compositions. The proof is based on the concept of a codeword chain. This result is then extended to the more general class of uniquely decipherable codes described by: $W_p = P_1 U P_2$,
where $P_1 = \{w_j \mid \lambda(w_j) = \ell,\ \ell > \lambda(w_i)\ \forall\, w_i \in P_2\}$ and P_2 is both a prefix and a suffix code.

The general case is then discussed and, using a theorem discovered by Block [6] and Carter and Gill [7], a conjecture on the replacement of uniquely decipherable codes by prefix codes with the same codeword compositions is proposed. Some implications of this conjecture are then discussed. No device analogous to a codeword chain could be found for the general case. Attempts to prove or disprove the conjecture have been unsuccesful so far.

BIBLIOGRAPHY

1. R. Ash, Information Theory, Interscience Publishers, New York, 1965, pp. 1-44.

2. R. M. Karp, "Minimum Redundancy Coding for Discrete Memoryless Channel," IRE Trans. Information Theory, Vol. IT-7, 1961, pp. 27-38.

3. D. A. Huffman, "A Method for the Construction of Minimum Redundancy Codes," Proc. IRE, Vol. 40, 1952, pp. 1098-1101.

4. L. G. Kraft, "A Device for Quantizing, Grouping and Coding Amplitude Modulated Pulses," M.S. Thesis, 1949, Electrical Engineering Dept., MIT, Cambridge, Mass.

5. B. McMillan, "Two Inequalities Implied by Unique Decipherability," IRE Trans. Information Theory, Vol. IT-w, 1956, pp. 115-116.

6. J. R. Block, "On Variable Length Codes with Prescribed Composition," M.S. Thesis, 1975, Electrical Engineering Dept., Northwestern University, Evanston, Ill.

7. L. Carter and J. Gill, "Conjectures on Uniquely Decipherable Codes," IEEE Trans. Information Theory, Vol. IT-20, 1974, pp. 394-396.

EFFICIENT INFORMATION FEEDBACK FOR BINARY ENCODED SIGNALS*

JOHN J. METZNER
Department of Electrical and Computer Engineering
Wayne State University
Detroit, Michigan 48202

ABSTRACT

An information feedback scheme is proposed whereby the sender tries to correct errors in an originally transmitted binary sequence by sending error-locating signals. Two different methods of transmitting the error locations are considered and analyzed, and achievable data rates are compared with ideal channel capacity.

I. INTRODUCTION

Schalkwijk and others[1-6] have considered information feedback schemes whereby one of M messages is sent (M large), and the receiver sends back information about the received signal; the transmitter thereafter sends corrective signals in order to lead the receiver to a correct decision. The general principle of leading the receiver to the correct decision also can be applied to binary encoded systems, as will be shown in this paper. The scheme is similar in philosophy to Schalkwijk's, but differs considerably in the details of its operation and performance. In Schalkwijk's procedure a numerical value is fed back and the transmitter sends a value linearly proportional to the difference between the receiver's numerical value and the exact numerical value which the transmitter wishes to convey. This process is repeated until the receiver's number becomes sufficiently close to the desired number. Kashyap[7] has extended this idea to vector quantities and the difference between a desired vector and the received vector. In the procedure to be described here, a binary sequence (which is presumed to contain some parity check digits) is sent, and the receiver sends back the received sequence after making binary decisions. The transmitter observes the errors and tries to iteratively reduce the number of errors by successive transmission of the locations of errors. The check digits are used simply for error detection (or limited error correction if practical) so that the receiver knows when an error-free situation has been reached.

The idea also is related to Chang's original idea of information feedback[8]. Sending back the exact binary sequence received corresponds to Chang's "total information feedback". However, Chang's scheme did not include an attempt to provide error location information. If the feedback signal did not check, Chang suggested that an erasure signal be sent and the message then be repeated. This "discarding information feedback" is less efficient than the scheme to be described, whereby, as with Schalkwijk's scheme, rates approaching capacity can be achieved as well as high reliability, yet without overly complex decoding requirements. Schalkwijk's scheme required a feedback information rate much greater than the forward rate, whereas in the present scheme the feedback channel rate is about the same as the forward rate. On the other hand, the present scheme is limited in rate by the binary channel capacity, which is lower than the general channel capacity which limits Schalkwijk's scheme.

By incorporating check digits, the present scheme provides a "fail-safe" protection[9] which prevents errors in the event of a temporary

---

*The research in this paper was sponsored by the National Science Foundation on Grant no. ENG 75-07839.

severe disturbance on the forward and/or feedback channel; the only affect is to cause a temporary interruption and start-over until conditions clear.

II. DESCRIPTION OF THE PROCEDURE

The first sending is a block of binary digits. Assume these are sent as binary orthogonal signals of effective power P, duration τ, in white gaussian noise of spectral density $N_o/2$. Let there be N digits in the block, of which K are data digits. The number N usually will be considered large (at least 500, and possibly much greater). The number N-k of check digits will be sufficient (say 30 or 40 in the pure error detection case) to provide for negligibly small probability of an undetected error. Thus K/N normally will be fairly close to 1. Channel capacity increases with bit rate. (μ binary digits/second), approaching a constant as $\mu \to \infty$. This is illustrated in figure 1 by the plot of normalized binary orthogonal channel capacity $C_B N_o/P$ vs. $\tau P/N_o$. ($\mu = \frac{1}{\tau}$). The limit as $\tau \to o$ (or $\mu \to \infty$) is $\frac{1}{\pi} \log_2 e$. (For antipodal binary it would be $\frac{2}{\pi} \log_2 e$, and without a binary limitation it would be $\log_2 e$). The corresponding digit error probabilities are indicated along a horizontal scale. It can be seen that higher efficiency calls for high bit rate and consequently acceptance of a rather large proportion of binary decision errors. Also, at high efficiencies, the capacity in bits per digit will be significantly less than one, and there will be virtually no chance that the first sending will be decodable with high reliability.

The first sending is returned to the transmitter via a feedback channel. The feedback channel is assumed to be almost noiseless - the effect of feedback errors usually will be to result in starting over (retransmission) due to failure to arrive at a decodable result after a specified number of iterations or after a specified time. Thus feedback errors result in a reduction in throughput rate but need not cause a deterioration in reliability. By "almost noiseless" is meant that the rate reduction due to feedback errors is not significant.

There must be a time delay between successive sendings of information about a given block due to feedback round trip time. However, since it is possible to interlace other sendings during these waiting times[10], these time gaps will be ignored in determining data rate in bits per second.

After the return of the first sending the sender observes a certain number, say e_1, of errors. The suggested procedure is for the transmitter to then send simultaneously e_1 out of N orthogonal signals of duration T_1 (which T_1 is to be optimized) and using energy PT_1/Np_1 per signal component, where p_1 is the per-digit probability of error on the first sending. The above is based on maintaining an average power P equal to that used for the original binary digits. The key quantity to performance is the energy per signal component. Various orthogonal signal sets could be used, but they all yield the same performance in white noise for the same energy per signal. For example, the orthogonal signals could be N non-overlapping waveforms within duration T_1/N each, with "spikes" of average power P/p_1 in the intervals corresponding to the error locations. Alternatively, they could be N different frequency sinusoids of duration T_1, the sum of e of them chosen corresponding to the error locations. The receiver then decides the locations of the errors by either a presence-absence decision on each orthogonal component or in some circumstances, by choosing the greatest e'_1 of N outputs. (In this latter case, it must be realized that the receiver does not know e_1. However, due to the error detection fail-safe provision, the receiver could in stages try to decode first by just changing the location with the greatest output, then, by just changing the two greatest location outputs, etc., until a decodable result were reached or until a certain number of changes were made).

Assume for now that a presence-absence decision will be made on each orthogonal component in the correctional signal. The errors remaining after the correctional signal will correspond precisely to the orthogonal components decided incorrectly in the second reception. Let p_2 be the probability of error for these decisions. The strategy will be such as to ensure that $p_2 < p_1$. If the sequence still is not decodable after the second sending, one possibility is for the receiver to return the updated binary decisions or, alternatively, to send back the e'_1 of N correction decisions in a manner similar to the forward technique. There will generally be a considerably smaller number e_2 of errors remaining. These can be corrected by sending e_2 out of N orthogonal signals of duration T_2, with energy PT_2/Np_2 per component. Since there is now usually considerably more power per component, T_2 usually can be made appreciably smaller than T_1, while actually increasing the energy per component. In other words,

$$\frac{P\,T_2}{Np_2} > \frac{P\,T_1}{Np_1}\ , \text{ but } T_2 < T_1\ .$$

This process can be continued until the word is decoded, with the proviso that if a certain number of iterations is reached the result is discarded and the original block is retransmitted. It will be seen that if the initial digit error probability is about .06 or less, just a single iteration (e_1 out of N only) is almost as good as an unlimited number of iterations. In general, it appears that at most two or three iterations are sufficient to yield most of the benefits of the scheme.

III. BASIC RATE AND CAPACITY RELATIONS

It is well known[11] that for equiprobable binary orthogonal signals of duration τ in white gaussian noise, with maximum likelihood detection,

$$P_B = \operatorname{erfc}_* \sqrt{\frac{\tau P}{N_o}} = \int_{\sqrt{\frac{\tau P}{N_o}}}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx \tag{1}$$

For comparison purposes, the capacity with restriction to binary orthogonal signals is

$$C_B = \frac{1}{\tau}\,[1 + P_B \log_2 P_B + (1-P_B)\log_2 (1-P_B)] \text{ bits/Second.} \tag{2}$$

as $\tau \to o$,

$$C_B \to \frac{1}{\pi}\frac{P}{N_o} \log_2 e\ , \tag{3}$$

where e is the natural logarithm base. Figure one includes a plot of $C_B N_o/P$ versus $\tau P/N_o$.

The duration of the original block is $N\tau$.

Let

$$T_i = \alpha_i N\tau, \tag{4}$$

and let T_f be the duration of the final iteration.

Then $N\tau[1 + \sum_{i=1}^{f} \alpha_i]$ is the total duration for sending K bits if the word can be decoded after the f iterations. Assume that the receiver always waits for the f iterations, and let P_R be the probability that retransmission then will be required. The effective rate R then is

$$R = \frac{1}{\tau[1 + \sum_{i-1}^{f} \alpha_i]} \frac{K}{N} (1-P_R). \tag{5}$$

Usually, $\frac{K}{N}(1-P_R)$ would be close to 1. Normalized,

$$\frac{R}{P/N_o} = \frac{1}{\frac{P\tau}{N_o}(1 + \sum_{i=1}^{f} \alpha_i)} \frac{K}{N} (1-P_R) \tag{6}$$

IV. RESULTS FOR PRESENCE-ABSENCE DECISIONS ON ERROR LOCATING SIGNALS

For the error-locating signal, let us assume for now that a present-absent maximum likelihood decision is made for each orthogonal waveform. Somewhat better performance could be attained by taking into account that "no error" is more likely than "error", but we will not consider this possible approach. Since the energy per component is PT_1/NP_1, the digit error probability p_2 on the second sending is

$$p_2 = \text{erfc}_* \sqrt{\frac{P\alpha_1 N\tau}{2NP_1 N_o}}$$

Let $$\lambda = \sqrt{\frac{P\tau}{N_o}} \tag{7}$$

Then $$p_1 = \text{erfc}_* \, \lambda \tag{8}$$

and $$p_2 = \text{erfc}_* \, \lambda \sqrt{\frac{\alpha_1}{2p_1}} \tag{9}$$

(Note that the factor of $\sqrt{2}$ appears in (9) but not in (8). This is because the first decisions are between two orthogonal signals, while subsequent decisions are presence-absence (or on-off) for each orthogonal component.) For a succession of correctional signals, we have the relation

$$p_{i+1} = \text{erfc}_* \, \lambda \sqrt{\frac{\alpha_i}{2p_i}} \tag{10}$$

The α_i should be chosen to minimize (5).

As an illustration, suppose that $N = 2000$, $K = 1950$ and $p_1 = 0.1$, so that 200 errors were expected initially. The locations of the $e_1 \cong 200$ errors are sent by a sum of orthogonal signals of relative duration $\alpha_1 = 0.68$ (68% of original block duration). Note that the correctional signal has energy $\frac{\alpha_1}{p_1} P\tau = 6.8P\tau$ per error-locating component, compared to $P\tau$ per original binary component. Thus the second sending is more reliable and, from (9), $p_2 = .0090$, and the expected number of errors is reduced from 200 to 18. Using $\alpha_2 = 0.2$ for a second correctional sending, with energy $\frac{\alpha_2}{p_2} P\tau = 22.2P\tau$ per error locating signal, the error probability is reduced to $p_3 = 10^{-5}$, and the expected number of errors per block is reduced to 0.02. The bound $P_R \leq .02$ is then a close one. From (6),

$$\frac{R}{P/N_o} \cong \frac{1}{(1.645)\ (1 + .68 + .2)}\ \frac{1950}{2000}\ (1-.02) = 0.309.$$

The above example is shown as point A in figure one.

The curve of $C_B N_o/P$ vs. $\tau P/N_o$ does not truly represent a bound on the rate of the orthogonal signal repeat scheme because the orthogonal error location signals are not assumed to be restricted to the same bandwidth and effective digit rate as the original binary sequence. (Since N binary orthogonal signals are sent in time $\alpha N\tau$, the net result is to increase bandwidth requirements by a factor $1/\alpha_{min}$ compared to the original τ second digits, where α_{min} is the smallest α_i for any of the repeats). However, the unlimited bandwidth upper bound of $\frac{1}{\pi}\log_2 e$ remains if a binary decision must be made on the presence or absence of each orthogonal signal. A closer bound can be obtained by the following reasoning. The remaining average uncertainty after the first sending is $Nh(p_i)$. If this is to be reduced to zero by a signal of duration $N\tau \sum_{i=1}^{f} \alpha_i$, we must have

$$\frac{1}{\pi}\frac{P}{N_o}\log_2 e \geq \frac{Nh(p_1)}{N\tau \sum_{i=1}^{f} \alpha_i} \tag{11}$$

or

$$\sum_{i=1}^{f} \alpha_i > \frac{\pi h(p_1)}{\frac{\tau P}{N_o}\log_2 e}\ . \tag{12}$$

For large N and small rejection probability after f iterations,

$$\frac{R_{EFF}}{P/N_o} = \frac{1}{P\tau/N_o(1+\sum_{i=1}^{f} \alpha_i)} \tag{13}$$

Substituting (12) in (13),

$$\frac{R_{EFF}}{P/N_o} \leq \frac{1}{P\tau/N_o + \frac{\pi h(p_1)}{\log_2 e}}\ . \tag{14}$$

This upper bound is plotted as bound curve X in figure one.

Included in figure one is a curve of rate R versus λ for the case where N = 2000, K = 1950, two error-locating signal blocks are sent, and the receiver uses error detection only on the block to determine if it is acceptable. At each λ, the pair α_1, α_2 is optimized. Also shown (dashed line) is a curve for similar assumptions except that the receiver can use single-error correction as well as error detection. This latter capability improves achievable rate somewhat because the digit error probability does not have to be brought down as low to yield a low rejection probability. In all cases, the α_1, α_2 pair was chosen to yield a rejection probability of 0.02 after two error locating signal blocks. Maximum rate is achieved when the initial digit error probability is about 0.15.

Higher normalized rates can be achieved by allowing a larger number of error-locating blocks. This was not analyzed in detail, but point B illustrates a near-optimum case when three error-locating blocks are allowed.

V. GREATEST e OUT OF N DECISIONS FOR THE CASE OF ONE ERROR-LOCATING BLOCK

If only one error-locating block is used, a greatest e out of N decision rule may be feasible, since the fail-safe protection allows successive changing of the most likely errors until a decodable result is obtained or some maximum number of changes have been made. Suppose there are e errors initially. The receiver observes N orthogonal signal outputs $y_1, y_2, \ldots . y_N$. Let $y_1, y_2, \ldots . y_e$ be the outputs in the e error positions, and let $Y_{e+1}, \ldots . y_N$ be the outputs inthe N-e non-error positions. The former group consists of signal plus gaussian noise, while the latter consists of noise alone. To correct all the errors with a greatest e out of N decision criterion, it is necessary that all $y_1, \ldots . y_e$ be greater than all $y_{e+1} \ldots . y_N$. An integral expression for the probability of this can readily be written, but a simple and fairly close bound can be obtained as follows: if all $y_1 \ldots . y_e >$ all $y_{e+1} \ldots . y_N$, then there are e(N-e) binary inequalities, $y_i > y_j$, $i = 1,2, \ldots e$, $j = e+1 \ldots N$, which must simultaneously be satisfied. The probability of each case individually being satisfied is just like the probability of correct decision with binary orthogonal signals. Thus, by the union bound for probabilities,

$$P\binom{\text{all e errors are}}{\text{not corrected}} \leq e(N-e) \operatorname{erfc}_* \lambda \sqrt{\frac{\alpha_1}{p_1}} \quad . \tag{15}$$

Equation (15) is for a particular value of e. Thus,

$$P_R \leq \sum_{e=1}^{N} \binom{N}{e} p_1^{\,e} (1-p_1)^{N-e} \; e(N-e) \operatorname{erfc}_* \lambda \sqrt{\frac{\alpha_1}{p_1}} \tag{16}$$

which reduces to

$$P_R \leq N(N-1)\, p_1 (1-p_1) \operatorname{erfc}_* \lambda \sqrt{\frac{\alpha_1}{p_1}} \tag{17}$$

Curve Y in figure one is computed for the case N = 600, K = 560, using equations (6) and (17), with the best α selected for each value of λ.

VI. AN ALTERNATIVE PROCEDURE

A different approach toward reducing the errors would be to send, one by one, short binary sequences, of length L each, where L is the smallest integer $\geq \log_2 N$, giving the location of one error each. These are returned to the transmitter as received. Each short sequence received results either in an increase or a decrease by one in the number of errors. Eventually the number of errors is reduced to zero or to a number that can be self-corrected at the receiver. The duration and power in the binary pulses of the error location signals may or may not be the same as the original binary pulses. The following analysis yields an approximation to the efficiency of this scheme. Both initial and error-locating pulses are assumed to have the same power P and duration .

The probability that the L-digit error-locating block is received incorrectly is,

$$P_L = 1 - (1-p_1)^L = 1 - Q_L \tag{18}$$

If, at any stage there are e errors in the message, the probability that the number of errors will increase to e+1 is $P_L\left(\frac{N-e}{N-1}\right)$, and the probability it will decrease to e-1 is $1 - P_L\left(\frac{N-e}{N-1}\right)$. Thus, a state diagram could be drawn to study the reduction of the number of errors with successive iterations. This is messy, however, and a good simple bound is obtained by assuming P_L (where $P_L \geq P_L\left(\frac{N-e}{n-1}\right)$ if $e > 0$) is the probability of adding an error. This converts the problem into a standard gambler's ruin problem, where, if $\overline{D}_B$ is the expected number of location signals to reduce the number of errors from e to e_o with the P_L bound assumption it is well known that if $e \geq e_o$,

$$\overline{D}_B = \frac{e - e_o}{Q_L - P_L} \tag{19}$$

(If $e \leq e_o$, then of course $\overline{D}_B = 0$.)
Thus, the true expected number of location signals $\overline{D}$ is bounded by this, or

$$\overline{D} \leq \frac{e - e_o}{Q_L - P_L}, \quad e \geq e_o. \tag{20}$$

Another possibility which is more efficient in some cases is to send one additional parity digit with each location signal. Then the number of errors can increase by one, remain the same, or decrease by one at each stage. For this case, $\overline{D}$ is again bounded by the expression (20), but now $Q_L \neq 1 - P_L$.

Note that $\overline{D}$ is averaged for a particular e. Averaging overall e,

$$\overline{\overline{D}} = \sum_{e \geq e_o} \frac{(e-e_o)\ P(e)}{Q_L - P_L} \tag{21}$$

The only cases we will consider in detail are $e_o = 0$ and $e_o = 1$. If $e_o = 0$,

$$\overline{\overline{D}} = \frac{Np_1}{Q_L - P_L} \tag{22}$$

If $e_o = 1$,

$$\overline{\overline{D}} = \frac{Np_1 - \text{Prob.}(e \geq 1)}{Q_L - P_L} \tag{23}$$

Since each block requires an average of $N + L\overline{\overline{D}}$ digits,

$$\frac{R}{P/_{N0}} = \frac{K}{\lambda^2\ (N + L\overline{\overline{D}})} \tag{24}$$

Equation (24) is plotted for K = 482, N = 512 and three different cases. One case is for a 9-bit uncoded error location sequence and $e_o = 0$ (eq. 22 for $\overline{D}$). The second and third cases are for a 10-bit error location sequence with one parity check digit included, both for $e_o = 0$ and $e_o = 1$. It is seen that the check digit improved performance somewhat. The ability to correct single errors in the block appears to be of little advantage.

VII. DISCUSSION

The orthogonal error location scheme using two error locating blocks permits reliable transmission at rates equivalent to about 1.5 db difference in signal power compared to unlimited bandwidth binary orthogonal channel capacity, while avoiding the need for complex error correction schemes. The system may require a bandwidth several times that of the original binary sequence, as well as a good feed-back channel of comparable bandwidth. A single stage of error location is somewhat less efficient (about 3 db difference in power compared to limiting binary orthogonal capacity). However, by operating at a digit error probability of about 0.06 instead of about 0.15, the bandwidth requirement is less (larger τ) for a given average power P. The scheme of locating errors one at a time is still less efficient, yet is close to ideal if the initial digit error probability is constrained to be about 0.03 or less. Also, this latter scheme does not increase the bandwidth beyond the original digit rate requirement.

Although we have emphasized error detection only at the receiver, there are some error patterns, such as bursts of errors or one or two random errors, which are relatively easy to correct, and this would result in some efficiency improvement. Also, the sender, knowing the receiver's decoding capabilities, could select which errors to locate. For example, if the receiver had burst error correcting capability and the error pattern was a burst plus a few scattered random errors, the sender could merely correctly locate the scattered random-errors, leaving the burst which the receiver could then correct. By sending fewer error locations, the sender could use less energy, or send more reliably with the same total energy.

REFERENCES

1. J.P.M. Schalkwijk and T. Kailath, "A Coding Scheme for Additive Noise Channels with Feedback," IEEE Transactions on Information Theory, vol. IT-12, no. 2, pp. 172-189, April 1966.

2. S.S. Lavenberg, "Feedback Communication using Orthogonal Signals," IEEE Transactions on Information Theory, vol.IT-15, no.4,p.478-483 July,1969.

3. J.P.M. Schalkwijk, "Center-of-Gravity Information Feedback," IEEE Transactions on Information Theory, vol.IT-14, pp.324-331, March 1968.

4. A.J. Kramer, "Improving Communication Reliability by use of an Intermittent Feedback Channel," IEEE Transactions on Information Theory, vol. 15, no. 1, pp. 52-60, January 1969.

5. J.F. Hayes, "Studies of Sequential Detection Systems with Uncertainty Feedback," IEEE Transactions on Information Theory, vol.IT-14, no. 1, pp. 63-70, January 1968.

6. Omura, J.K., "Optimum Linear Transmission of Analog Data for Channels with Feedback,"IEEE Transactions on Information Theory, vol. IT-14, no. 1, pp. 38-43, January 1968.

7. Kashyap, R.L., "Feedback Coding Schemes for an Additive Noise Channel with a Noisy Feedback Link," IEEE Transactions on Information Theory, vol. IT-14, no. 3, pp. 471-480, May 1968.

8. Chang, S.S.L., "Theory of Information Feedback Systems," IRE Transactions on Information Theory, vol.IT-2, no.3, pp.29-40, September 1956.

9. J.J. Metzner and K.C. Morgan, "Reliable Fail-Safe Binary Communication" IRE Wescon Convention Record, vol. 4, pt. 5, pp.192-206, 1960.

10. J.J. Metzner and K.C. Morgan, "Word Selection Procedures for 'Error-Free' Communication Systems," New York University, Second Scientific Report on Contract AF 19(628)-4321, June 1965.

11. Van Trees, H.L., Detection Estimation and Modulation Theory, Part I, John Wiley & Sons, 1968.

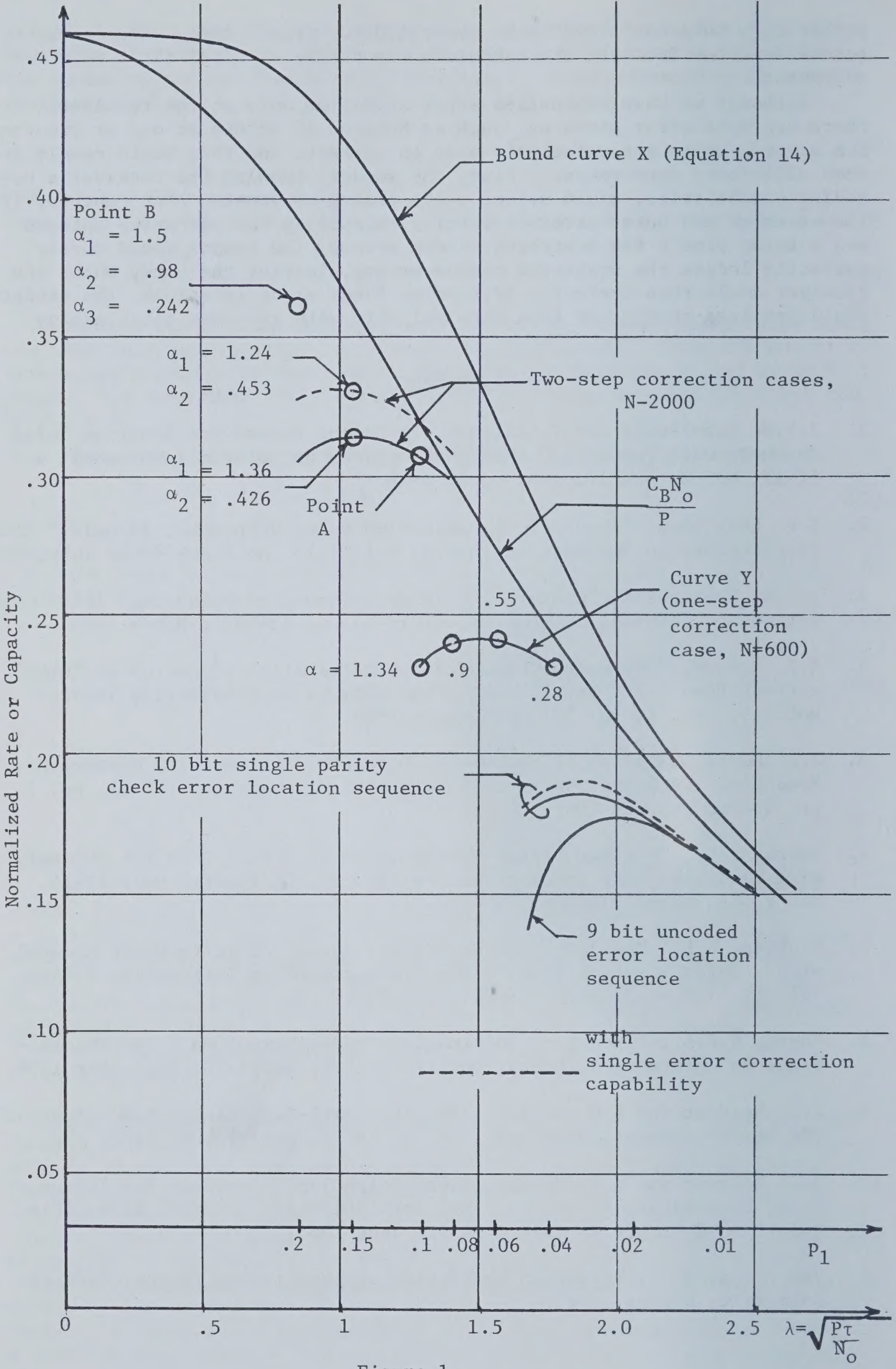

Normalized Rate or Capacity
.45
.40
.35
.30
.25
.20
.15
.10
.05
Bound curve X (Equation 14)
Point B
$\alpha_1 = 1.5$
$\alpha_2 = .98$
$\alpha_3 = .242$
$\alpha_1 = 1.24$
$\alpha_2 = .453$
Two-step correction cases, N-2000
$\alpha_1 = 1.36$
$\alpha_2 = .426$
Point A
$\frac{C_B N_o}{P}$
.55
Curve Y (one-step correction case, N=600)
$\alpha = 1.34$
.9
.28
10 bit single parity check error location sequence
9 bit uncoded error location sequence
with single error correction capability
.2
.15
.1
.08
.06
.04
.02
.01
P_1
0
.5
1
1.5
2.0
2.5
$\lambda = \sqrt{\frac{P\tau}{N_o}}$

Figure 1

A THEOREM ON CYCLIC CODES WITH APPLICATIONS[*]

GERALD E. SEGUIN
Department of Electrical Engineering
Royal Military College of Canada
Kingston, Ontario, Canada

&

VIJAY K. BHARGAVA[**]
Department of Electrical Eng.
Concordia University
Montreal, Quebec, Canada

SUMMARY

We prove the following theorem.

Theorem: Let V be a binary (n, k) cyclic code with generator polynomial $g(x)$ and where $n = mk$. Then $g(x), x^{m}g(x), \ldots, x^{m(k-1)}g(x)$ is a basis for V if, and only if, for every factor $f(\lambda)$ of $\lambda^{k}-1$, $f(\lambda) \neq \lambda^{k}-1$, $h(\lambda)$ does not divide $f(\lambda^{m})$; where $g(\lambda)\, h(\lambda) = \lambda^{n}-1$.

Applications: Using the above theorem we obtain a number of (mk, k) cyclic codes into systematic quasi-cyclic form, i.e., codes with generator matrices of the form

$$G = [I_k, \; C_1, \; C_2, \; \ldots, \; C_{m-1}]$$

where each C_i is a mxm circulant matrix. Some of the advantages of having the code in a systematic quasi-cyclic form are: (1) Ease in computing the weight distribution. (2) Easy encoding and decoding procedures using the methods proposed by Maurice Karlin. (3) The code as well as the dual code are specified by the same circulants.

---

[*] This work was supported in part by the National Research Council of Canada, under grants A-7382 and A-7787.

[**] The work of this author was done while he was with the University of Waterloo, Waterloo, and with Queen's University, Kingston, Canada.

PERFORMANCE ANALYSIS OF BINARY SEQUENTIAL COMMUNICATION SYSTEMS EMPLOYING HARD DECISION STRATEGY

ENZO DALLE MESE* and DINO GIULI**
*Istituto di Elettronica, University of Pisa, Italy
** Istituto di Elettronica, University of Florence, Italy

In this paper we analyse the performance which can be achieved by employing a sequential decision strategy for the binary symmetric channel (BSC). The receiver performs a probability-ratio-sequential-test on the received bit stream which corresponds to the information bit generated by the source and sent repetitively by the transmitter. An analysis of the asympthotic behavior of the resulting scheme allows us to single out the class of communication systems for which the proposed sequential scheme is asympthotically better than the corresponding one-way system. In particular the coherent PSK modulation scheme with additive white Gaussian noise channel (AWGNC) and quantized receiver belongs to that class; that does not happen in the case of FSK modulation scheme with noncoherent AWGNC and quantized receiver.

INTRODUCTION

Recent literature has paid considerable attention to feedback communication systems (see for-example [1,2] and references therein). In this paper binary communication systems with time-quantization (i.e. in which the decision is made at integer bit periods) are considered, since these systems prove simpler to analyze and to implement.

The structure of the optimal receiver in these systems is well known [1,3]. It can be considered a soft decision receiver [4] since the final decision is taken on the basis of the output samples of an unquantized demodulator (Fig. 1a).

In this paper we analyse the performance of a suboptimal system which uses a hard decision receiver, obtained by carrying out a quantized demodulation (Fig. 1b). The optimal sequential strategy and the system performance are determined under the hypothesis that the forward channel between points A and B in Fig. 1b can be considered a BSC. In particular, the optimal asympthotic efficiency of the system is calculated. This allows us to single out the class of communication systems for which the proposed sequential scheme is potentially efficient.

Finally we calculate in detail the performance obtained by applying the proposed sequential procedure both in the case of PSK modulation-AWGNC and of FSK modulation-noncoherent AWGNC.

OPTIMAL SEQUENTIAL RECEIVER FOR BSC

In the analysis of the system in Fig. 1b the following hypotheses are made:

i) the symbols $\{m_{\pm 1}\}$, generated by the source, are equally likely and independent;

ii) the transmitted signals $s_{+1}(t)$ and $s_{-1}(t)$ which correspond to m_{+1} and m_{-1} respectively, are defined on the time interval $(0,T)$;

iii) the feedback channel is noiseless and delayless.

Under the further hypothesis that the forward channel between points A and B in Fig. 1b can be considered a memoryless BSC with error probability ϵ, the analysis of the system in Fig. 1b is equivalent to the analysis of the system in Fig. 2. The problem then consists in determining the optimal strategy which the sequential decision device is to follow.

As is well known [5], the optimal sequential strategy consists in comparing at each step the logarithm of the likelyhood ratio with two suitable thresholds. Under the above assumptions the optimal sequential strategy becomes [6,7]:

$$\begin{cases} \text{if } d_n = S & \text{choose } m_{+1} \\ \text{if } d_n = -S & \text{choose } m_{-1} \\ \text{if } -S < d_n < S & \text{make another observation} \end{cases} \tag{1}$$

where

$$d_n \triangleq 2m_n - n \tag{2}$$

$$m_n \triangleq \text{number of +1's present in the first n bit received} \tag{3}$$

and S is a discrete threshold depending on the overall error probability P_e.

The strategy synthesised in (1) leads to a simple realization of the sequential decision device. In fact, this can be realized with a binary counter which either increases or decreases by one depending on whether +1 or −1 is received, and which is reset whenever its contents reach one of two set digital thresholds.

The parameters which define the performance of the system in Fig. 2 are P_e, the error probability, and $\overline{N}$, the average number of repetitions of the same symbol needed before making a decision. It follows that $(1/\overline{N}T)$ is the transmission rate. By observing that the random variable d_n defined in (2) follows a random walk between thresholds −S and S with transitional probabilities ϵ and $1-\epsilon$, it is easy to obtain [6,7] :

$$P_e = \frac{1}{1+\left(\frac{1-\epsilon}{\epsilon}\right)^S} \tag{4}$$

$$\overline{N} = \frac{S}{1-2\epsilon} \, \frac{1-[\epsilon/(1-\epsilon)]^S}{1+[\epsilon/(1-\epsilon)]^S} \tag{5}$$

As could be foreseen, since $0 < \epsilon < 1/2$, from eq. (4) and (5) we deduce, that, for a fixed ϵ, P_e can be made as small as one wishes by increasing S and therefore $\overline{N}$. This merely means that the decrease in P_e is obtained by increasing the average energy transmitted per symbol. In practice, to calculate the efficiency of a system like the one shown in Fig. 2 it is necessary to determine the dependence of P_e on $\overline{N}$ with a fixed average transmitted energy. For this it is necessary to know the dependence of ϵ on the energy of the transmitted signals.

ASYMPTHOTIC EFFICIENCY OF THE SYSTEM

To determine the performance of a specific system it is certainly useful to know its asympthotic efficiency. For this, supposing that the transmitted signals both have energy E, we determine, for the system in Fig. 2, the limit $P_{e\infty}$ of P_e as $\overline{N} \to \infty$, mantaining a constant average transmitted energy $\overline{N}E$. Since $\epsilon \to 1/2$ as $E \to 0$, in a neighbourhood of E=0 we can write

$$\epsilon = (1/2)+O(E) \tag{6}$$

where $O(E) \to 0$ as $E \to 0$. From eq. (4) and (5) we obtain

$$\overline{N}(1-2\epsilon)\log \frac{1-\epsilon}{\epsilon} = (1-2P_e)\log \frac{1-P_e}{P_e} \tag{7}$$

Since, in a neighbourhood of E=0,

$$\log \frac{1-\epsilon}{\epsilon} = \frac{1-2\epsilon}{\epsilon} + \text{h.o.t.} \tag{8}$$

we have from eq. (7)

$$\lim_{\substack{\overline{N}\to\infty \\ E\overline{N}=\text{const.}}}(1-2P_e)\log \frac{1-P_e}{P_e} = \lim_{\substack{\overline{N}\to\infty \\ E\overline{N}=\text{const.}}} \overline{N}\,\frac{(1-2\epsilon)^2}{\epsilon} \tag{9}$$

Supposing that in (6)

$$O(E)=kE^{\alpha} + \text{h.o.t.} \tag{10}$$

from (9) we have

$$\lim_{\substack{\overline{N}\to\infty \\ E\overline{N}=\text{const.}}} (1-2P_e)\log \frac{1-P_e}{P_e} = \lim_{\substack{\overline{N}\to\infty \\ E\overline{N}=\text{const.}}} 8k^2\overline{N}E^{2\alpha} = \lim_{E\to 0} k'E^{2\alpha-1} \tag{11}$$

where the constant

$$k'=8k^2\overline{N}E > 0 \tag{12}$$

The limit in (11) depends on the value of α ; in fact, we have the following cases:

a) for $\alpha > 1/2$:

$$\lim_{\substack{\overline{N}\to\infty \\ E\overline{N}=\text{const.}}}(1-2P_e)\log \frac{1-P_e}{P_e} = 0$$

which implies

$$P_{e\infty} = 1/2 \tag{13}$$

b) for $\alpha = 1/2$:

$$\lim_{\substack{\bar{N}\to\infty \\ E\bar{N}=\text{const.}}} (1-2P_e)\log\frac{1-P_e}{P_e} = k'$$

which yields in implicit form the value of $P_{e\infty}$. Since $P_{e\infty}$ is a decreasing function as $\bar{N}E$ increases, we can obtain an approximate explicit expression of $P_{e\infty}$ for sufficiently large values of $\bar{N}E$, and corresponding small values of $P_{e\infty}$ ($P_{e\infty} < 10^{-2}$):

$$P_{e\infty} \cong \exp(-k') = \exp(-8k^2\bar{N}E) \tag{14}$$

c) for $\alpha < 1/2$:

$$\lim_{\substack{\bar{N}\to\infty \\ E\bar{N}=\text{const.}}} (1-2P_e)\log\frac{1-P_e}{P_e} = \infty$$

which implies

$$P_{e\infty} = 0 \tag{15}$$

The analysis carried out shows that only in cases b) and c) we can expect a satisfactory improvement in the performance of the system in Fig. 1b compared to the corresponding one-way system.

APPLICATIONS

In this section we will examine the convenience of the sequential strategy proposed compared to the non-sequential strategy for two practical binary communication systems, bearing in mind the results in the preceding section.

Coherent PSK modulation-AWGNC

In this case the demodulator in Fig. 1 is a filter matched to one of the two transmitted signals, followed by a sampler, while the two-level quantizer has a zero threshold. This choice which, as is well known, optimizes the corresponding one-way system (1) satisfies the hypotheses in the second section. The resulting BSC is characterized by an error probability

$$\epsilon = \frac{1}{2} - \frac{1}{\sqrt{2\pi}}\int_0^{\sqrt{E/N_0}} \exp(-x^2/2)\,dx \tag{16}$$

where N_0 is the power spectral density of the noise. In this case, by expanding in power series, we obtain

$$O(E) = -[E/(2\pi N_0)]^{\frac{1}{2}} + \text{h.o.t.}$$

and therefore

$$\alpha = 1/2 \quad , \quad k = -1/\sqrt{2\pi N_0} \tag{17}$$

---

(1) Also note that this choice allows the transformation of the continuous-time system into a discrete-time system without any loss of information, since the sequence $\{z_n, n=1,2,\ldots\}$ is a sufficient statistic for estimating the transmitted symbols.

This system then falls in case b) of the preceding section with

$$P_{e\infty} \cong \exp(-\frac{4}{\pi}\,\overline{N}\,\frac{E}{N_o}) \qquad (18)$$

Observe that $P_{e\infty}$ is always less than the error probability of the corresponding optimum one-way system, so that we can expect an improvement in the use of the sequential strategy. The performance of the sequential system is shown in Fig. 3 (continuous curves) compared with those of the optimum one-way system (curve corresponding to S=1). In the same figure the curve giving $P_{e\infty}$ of the optimum unquantized two-way system [3] is reported for comparison purpose (dashed curve).

From Fig. 3 we can conclude:

i) the system proposed yields a consistent gain (2÷3 dB) compared to the one-way system in spite of the quantization employed;

ii) due to the quantization employed, there is a power loss (2÷3 dB) with respect to the non-quantized two-way system, which is nevertheless compensated for by the extreme simplicity of the resulting scheme.

FSK modulation-noncoherent AWGNC

In this case the demodulator in Fig. 1 is realized with two filters matched to the transmitted signals followed by two envelope detectors; the envelope detector outputs are sampled and then subtracted the one from the other. The resulting sequence $\{z_n\}$ is quantized with a zero threshold quantizer. This scheme which, as is well known, optimizes the one-way system satisfies the hypotheses in the second section; the resulting BSC is characterised by an error probability

$$\epsilon = (1/2)\exp(-E/2N_o) \qquad (19)$$

By expanding (19) in power series we have

$$O(E) = -E/(4N_o) + \text{h.o.t.}$$

and therefore

$$\alpha = 1 \quad ; \quad k = -1/(4N_o) \qquad (20)$$

This system then falls in case a) of the preceding section with $P_{e\infty} = 1/2$. The performance of this sequential system is shown in Fig. 4, with the same parameters as in Fig. 3. As we can see in Fig. 4, for $\overline{N}E/N_o > 10$ dB, we have a small improvement with respect to the one-way system for low values of S; anyway, the performance worsens rapidily with the increase of S, and therefore of $\overline{N}$, accordingly to (13).

REFERENCES

1. R.W.Lucky: "A Survey of the Communication Theory Literature: 1968-1973" IEEE Trans. on Inf.Theory, vol. IT-19, n. 5, 1973, pp. 725-739.

2. G.Prati, F.Russo, L.Verrazzani: "Programmed-Amplitude Binary Signaling with Decision and Information Feedback" IEEE Trans. on Aerospace and Electronic Systems, vol. AES-12, n. 2, 1976, pp. 225-228.
3. E.Dalle Mese: "Efficienza di un sistema di comunicazione binario coerente utilizzante reazione a decisione presa" Alta Frequenza, vol. XLII, n. 12, 1973, pp. 666-673.
4. Lin-Nam Lee: "On Optimal Soft-Decision Demodulation" IEEE Trans. on Inf. Theory, vol. IT-22, n. 4, 1976, pp.437-444.
5. A.Wald: "Sequential Analysis" J.Wiley & Sons, New York, 1947.
6. G.de Ruvo: "Sistemi di comunicazione con canale di reazione a decisione presa ad istanti discreti" Tesi di Laurea in Ingegneria Elettronica, Istituto di Elettronica, Università di Pisa, Aprile 1976.
7. E.Dalle Mese: "A new Decision Strategy in Binary Coherent Sequential Communication Systems" Int.Conf. on Inf. Sciences and Systems, August 19-24, 1976, Patras, Greece.

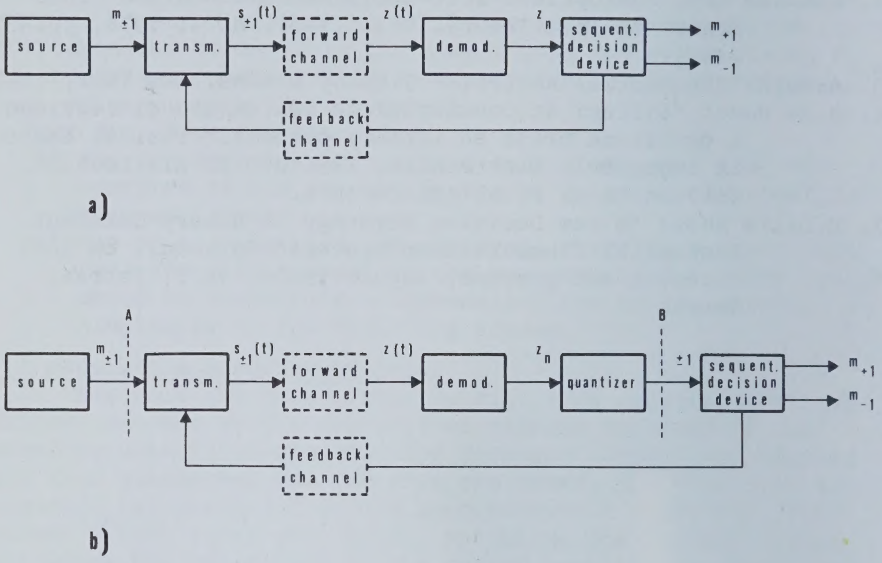

source
m±1
transm.
s±1(t)
forward channel
z(t)
demod.
zn
sequent. decision device
m+1
m-1
feedback channel
a)
A
source
m±1
transm.
s±1(t)
forward channel
z(t)
demod.
zn
quantizer
B
±1
sequent. decision device
m+1
m-1
feedback channel
b)

Fig. 1

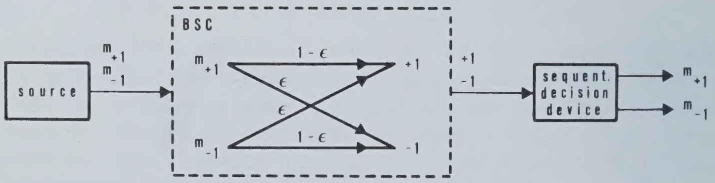
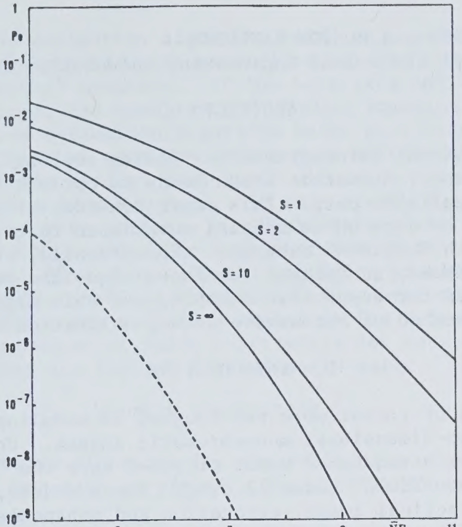
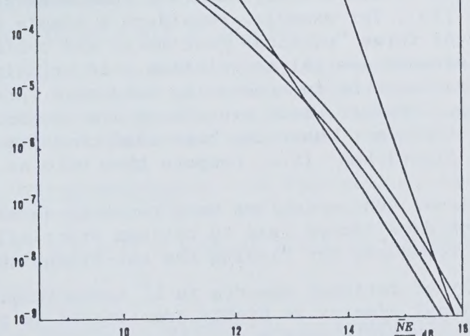

source
m+1
m-1
BSC
m+1
1 - ϵ
ϵ
ϵ
1 - ϵ
m-1
+1
-1
+1
-1
sequent. decision device
m+1
m-1

Fig. 2



1
Pe
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
S = 1
S = 2
S = 10
S = ∞
5
7
9
11
$\frac{\bar{N}E}{N_0}$ dB

Fig.3



1
Pe
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
S = ∞
S = 10
S = 5
S = 2
S = 1
10
12
14
16
$\frac{\bar{N}E}{N_0}$ dB

Fig.4

OPTIMUM LINEAR IMAGE RESTORATION OF REALISTIC DATA BASED ON THE TWO-DIMENSIONAL ARMA MODEL

HLUNG-WEN PENM and MORTON KANEFSKY
Department of Electrical Engineering University of Pittsburgh

ABSTRACT

A two-dimensional Autoregressive - Moving average (ARMA) model has beer recently developed by Strintzis which leads to optimum recursive enhancement procedures for realistic data. This paper considers the application of these procedures to data whose spatial covariance function appears to vary exponentially with Euclidean distance. Specifically, the identification problem is considered, an optimal recursive algorithm based on a simple ARMA model is developed for a specific example, and this algorithm is compared with the ad-hoc method of successive orthogonalization approximations.

1. INTRODUCTION

A great deal of recent work has focused on modeling restoration and enhancement of two-dimensional monochromatic images. Until recently, the most comprehensive statistical model proposed were the "Markov Mesh" model [1] and the "Markov Field" model [2,3,4]. These models, however, have not yet let to any practical image restoration and enhancement procedures. Because of this difficulty, models based on restrictive spatial correlation functions which fit known estimation procedures, the Kalman-Bucy linear recursive estimation procedure, in particular, have been considered. In these models, the correlation function ascribed to the picture is either one-dimensional stationary or two-dimensional stationary and factorable into the product of the respective point correlations along each of the two principal axis [5,6,7]. These assumptions are of highly questionable validity. In fact, it would be reasonable to assume, as in [2], that the correlation is a function of the Euclidean distance between points on the plane. Such a correlation function is not factorable. However, it was anticipated that these models would lead to simple recursive algorithms that require very little on-line storage. This has turned only to be incorrect [8,9], and these models no longer seem useful tools on which to base image restoration procedures.

Because of these modeling problems, suboptimal procedures based on arbitrary spatial correlation functions have been considered. The method of Feick & Strintzis [10], for example, considers a simple recursive algorithm based on the nearest three "previous" estimates and optimizes the parameters according to the assumed spatial correlation. In principle, one could approach an optimum estimate by incorporating many more "previous" estimates into the algorithm. However, such procedures are exceedingly complex computationally and, therefore, have not been used to document the suboptimality of the simple algorithms (i.e. compare them with an optimal estimation).

A two-dimensional ARMA model has been recently developed by Strintzis [11, 12, 15], which does indeed lead to optimum practical image enhancement procedures. The procedures for finding the one-dimensional ARMA model, are based on the family of rational spectra in L^2 space [13]. Unfortunately, the family of rational spectra is highly complicated in two dimensions [14] and does not lead to tenable ARMA models that are adequate for the representation of bivariate processes [15]. However, by using a different spectral model, Strintzis has developed a general ARMA model for bivariate processes which is entirely adequate for the formulation of simple optimum recursive estimation policies. The resultant algorithms are based on a limited number of previous estimates although more than the three estimates of Feick & Strintzis.

To implement the estimation algorithm based on the ARMA model, it is most convenient to characterize the spatial correlation by a two-dimensional form of the "Yule-Walker" equation. If the bivariate correlation function is known in closed form, the resultant Yule-Walker equation could be complex enough to result in an estimation algorithm based on a sizeable number of previous estimates, perhaps even the entire past history of the picture. However, the typical real-world situation involves estimating the spatial correlation from raw data. In this event, it is more natural and computationally easy to estimate the "Yule-Walker" relationship, which is adequate for the ARMA model, than any other formulation of the spectra or correlation function. This aspect of the modeling problem is considered in this paper. The results enable a comparison of recursive estimation algorithms based on the ARMA model with those based on the suboptimal successive orthogonalization procedure of Feick & Strintzis for data with spatial covariance functions that are typical of real-world data.

II. PROBLEM FORMULATION

We seek a recursive algorithm for the estimation of a zero-mean, bivariate process $x(m,n)$ embedded in $y(m,n)$. The observation noise $v(m,n)$ defined by

$$y(m,n) = x(m,n) + v(m,n) \tag{1}$$

is assumed to be additive, white, with power V, and independent of $x(m,n)$. The autocorrelation matrix of $x(m,n)$ is assumed to be stationary and known. The natural "past" of each observation $y(m,n)$ consists of all $y(i,j)$ with (i,j) in the Set $S(m,n)$ defined below and depicted in Figure 1.

$$S(m,n) = \{(i,j) : 0 \leq i \leq m-1)\ 0 \leq j\} \cup \{(i,j) : i = m,\ 0 \leq j \leq n-1\} \tag{2}$$

The optimum, mean-square, one-step prediction of $y(m,n)$ is denoted by $\hat{y}(m,n)$:

$$\hat{y}(m,n) = E\ \{y(m,n)/y(i,j),\ (i,j)\ \varepsilon\ S(m,n)\} \tag{3}$$

The optimum, mean-square, filtered estimate of $x(m,n)$ given both past and present observations will be denoted by $\hat{x}(m,n)$:

$$\hat{x}(m,n) = E\ \{x(m,n)/y(m,n) \cup y(i,j),\ (i,j)\ \varepsilon\ S(m,n)\} \tag{4}$$

The form of linear recursive algorithms reduce to

$$\hat{x}(m,n) = a(m,n)\ \hat{y}(m,n) + b(m,n)\ y(m,n). \tag{5}$$

We wish to consider algorithms where $\hat{y}(m,n)$ is based on a limited portion of the past and we will compare only the steady state version of these algorithms (i.e. determine and consider only the steady state values of $a(m,n)$, $b(m,n)$ and the estimation parameters of $\hat{y}(m,n)$).

a) Successive Orthogonalization. In this method, one arbitrarily assumes the form of the estimate $\hat{x}(m,n)$ to be

$$\hat{x}(k,s) = a_1(k,s)\ \hat{x}(k,s-1) + a_2(k,s)\ \hat{x}(k-1,s) + a_3(k,s)\ \hat{x}(k-1,s-1)$$
$$+\ b(k,s)\ y(k,s). \tag{6}$$

The coefficients are picked to satisfy the orthogonality principle:

$$E\{[\hat{x}(k,s) - x(k,s)] \begin{bmatrix} \hat{x}(k,s-1) \\ \hat{x}(k-1,s) \\ \hat{x}(k-1,s-1) \\ y(k,s) \end{bmatrix} \} = 0 \tag{7}$$

Equation (7) defines 4 equations from which the parameters of equation (6) can be determined. By successive solutions of these equations steady state values of the parameters can be found. The last of these 4 equations can be rewritten by incorporating equation (1), as

$$E\{[x(k,s) - \hat{x}(k,s)]\; x(k,s)\} = bV, \tag{8}$$

where b is the steady state solution of b(m,n). Thus, the minimum mean square steady state error is bV and the noise reduction performance of the filter in decibels is defined as

$$N = 10 \log_{10}(1/b) \text{ db}. \tag{9}$$

b) ARMA modeling. As in [13], one-dimensional ARMA models always specify a factorization of the two-sided z-transform rational function $\emptyset_o(z)$ of the autocovariance matrix $R_y(\tau) = E\{y(t + \tau)\; y(t)\}$. But in the two-dimensional case such a factorization may not exist for the rational function $\emptyset_o(z_1,z_2)$. Strintzis assumes the following type of two-dimensional z-transform of $R_y(i,j)$

$$\emptyset_1(z_1,z_2) \triangleq \sum_{i=0}^{+\infty} \sum_{j=-\infty}^{+\infty} R_y(i,j)\; z_1^i z_2^j \tag{10}$$

It is further assumed that $\emptyset_1(z_1,z_2)$ is an irreducible rational function of z_1 and z_2:

$$\emptyset_1(z_1,z_2) = \Big(\sum_{i=0}^{N_1} \sum_{j=0}^{N_2} a_{ij} z_1^i z_2^j\Big)^{-1} \Big(\sum_{i=0}^{N_1} \sum_{j=0}^{N_2} f_{ij} z_1^i z_2^j\Big) \tag{11}$$

with $a_{oo} = 1$. On combining these last two equations and equating coefficients of identical powers of z_1 and z_2 one obtains

$$\sum_{r=0}^{N_1} \sum_{s=0}^{N_2} a_{rs}\; u(t_1-r)\; R_y(t_1-r,t_2-s) = 0 \tag{12}$$

for all $(t_1,t_2) \notin \{(t_1,t_2):\; 0 \leq t_1 \leq N_1,\; 0 \leq t_2 \leq N_2\}$, where the unit step function u(x) equals one if x is nonnegative, zero if x is negative. One now defines, for $n > N_2$

$$w(m,n) \triangleq \sum_{r=0}^{N_1} \sum_{s=0}^{N_2} a_{rs}\; u(m-r)\; y(m-r,n-s). \tag{13}$$

With this formulation one can see that

$$\hat{y}(m,n) = -\sum_{\substack{i=0 \\ i+j \neq 0}}^{N_1} \sum_{j=0}^{N_2} a_{ij}\, y(m-i,n-j) + \hat{w}(m,n). \tag{14}$$

This is a convenient formulation because R_w and R_{wy} are non-zero in a limited region and $\hat{w}(m,n)$ becomes relatively easy to determine. It has been shown that [15]:

$$\hat{w}(m,n) = \sum_{T(m,n)} \frac{R_{w\varepsilon}(m,i;\ n,j)}{R_{w\varepsilon}(i,i;\ j,j)}[y(i,j) - \hat{y}(i,j)], \tag{15}$$

where $\varepsilon(i,j)$ is the error $[y(i,j) - \hat{y}(i,j)]$ or equivalently $[w(i,j) - \hat{w}(i,j)]$. The cross-correlation between $w(m,n)$ and $\varepsilon(p,q)$ can be found from

$$R_{w\varepsilon}(m,p;\ n,q) = R_{w}(m,p;\ n,q) - \sum_{T(m,n)\cap T(p,q)} \frac{R_{w\varepsilon}(m,i;\ n,j)}{R_{w\varepsilon}(i,i;\ j,j)}\ R_{w\varepsilon}{}^{T}(p,i;\ q,j) \tag{16}$$

for $T(m,n) = S(m,n) - \forall(m,n)$, where (17)

$$\forall(m,n) = \{(i,j) : 0 \leq i < m-N_1, 0 \leq j\} \cup \{(i,j) : i = m-N_1,\ 0 \leq j < n-N_2\} \tag{18}$$

and is shown in Figure 2. $R(m,p;\ n,q)$ is readily determined from equation (13). The mean square error $e(i,j)$ is given by

$$e(i,j) = R_{w\varepsilon}(i,i;\ j,j) = R_{y\varepsilon}(i,i;\ j,j). \tag{19}$$

Thus, equations (14-19) defines a recursive algorithm based on previous estimates and values limited to $T(m,n)$. This algorithm converges rather rapidly enabling one to find the steady state weight of $\hat{w}(m,n)$ and the steady state mean square error $e(i,j)$. Usually the number of "significant" weights are far fewer than the number of values in $T(m,n)$ resulting in a very near optimum algorithm that is easily implemented.

III. THE IDENTIFICATION PROBLEM

The intention is to compare the various image smoothing algorithms based on the reasonable assumption of a "Euclidean-distance" spatial correlation function. For no reason other than it appeared in a few articles [10, 16], the particular correlation function considered is

$$R_x(i,j) = 0.81 \exp\{-\sqrt{(0.44i)^2 + (0.34j)^2}\} \tag{20}$$

$$R_y(i,j) = R_x(i,j) + V\ \delta_{ij} \text{ where } \delta_{ij} = 1 \text{ if } i = j = 0, \text{ otherwise } \delta_{ij} = 0.$$

No effort was made to obtain an infinite expansion ARMA model for two reasons. First, a truncated version of this infinite expansion model would not be the best fit truncated ARMA model. Second, a procedure for determining approximate truncated ARMA models from a finite set of values of $R_x(i,j)$ is essential when working with raw data where a closed form expression for $R_x(i,j)$ is not known. The problem then becomes to "solve" the two-dimensional "Yule-Walker" equation (equation (12)), which means to find the "best-fit" coefficients of equation (12) for an arbitrary set of values for (N_1,N_2). Thus, for example, the (1,1) ARMA model becomes

$$R_y(i,j) = a_{10}R_y(i-1,j) + a_{01}R_y(i,j-1) + a_{11}R_y(i-1;\ j-1) \tag{21}$$

for all $(i,j) \notin \forall(1,1)$. The values of $R(i,j)$ in the region $\forall(1,1)$ become the necessary initial values.

We began the study by selecting arbitrary sets of 3 points outside of $\forall(1,1)$, and then determining the 3 a_{rs} weights from equation (20) and (21). The figure of merit for those solutions that was used was

$$\Sigma\Sigma[\hat{R}_y(i,j) - R_y(i,j)]^2, \tag{22}$$

where $\hat{R}_y(i,j)$ is the estimated value based on equation (21) and the sum was over a 50 x 50 array of points. The points which gave the best fit, according to equation (22), are indicated in Figure 3. This result is not particularly surprising. For one thing, we anticipated, by analogy to one-dimensional "Yule-Walker" identification, that points very close to $\forall(i,j)$ would be best since this would mean a precise fit where the values of $R(i,j)$ are largest and would be less susceptable to computer round-off error. The three points picked in this symmetrical fashion does appear to make them "closest" in some sense. Of significant interest to the identification problem is the fact that using these three points to determine the model requires the fewest number of estimates (9). For these reasons, this logical extension of this procedure for arbitrary (N_1,N_2) ARMA models seems most reasonable. Thus, in general, if c is the number of equations to be solved and d is the number of estimated values that are needed, then

$$c = (N_1 + 1)(N_2 + 1) - 1, \tag{23}$$

and

$$d = (2N_1 + 1)(2N_2 + 1). \tag{24}$$

Thus, d-c of these estimates are retained as initial values and the remaining c estimates are used to solve for the coefficients of the (N_1, N_2) truncated ARMA model. As another example, Figure 4 indicates the 25 estimates needed to identify the (2,2) ARMA model.

By assuming 9 measurements of the covariance function of data that is actually modeled by equation 20, one can determine the (1,1) ARMA model that fits these points. This model is from equation 12, for $V = 0.81/0.83$,

$$-R(m,n) = -0.776\, R(m,n-1) - 0.506\, R(m-1,n) + 0.346\, R(m-1,n-1),$$

$$\text{with } (m,n) \notin \{(m,n) : 0 \leq m \leq 1,\ 0 \leq n \leq 1\}. \tag{25}$$

The borders of the array are assumed to satisfy

$$-R(m,0) = -0.506\, R(m-1,0) \text{ for } m \geq 2, \tag{26}$$

$$-R(0,n) = -0.776\, R(0,n-1) \text{ for } n \geq 2, \tag{27}$$

for the purposes of determining the steady state algorithm.

Similarly if 25 measurements are assumed, the (2,2) ARMA models that fits these points become:

$$-R(m,n) = -0.926\, R(m,n-1) + 0.217\, R(m,n-2) - 0.713\, R(m-1,n) + 0.481\, R(m-1,n-1) - 0.070\, R(m-1,n-2) + 0.106\, R(m-2,n) - 0.025\, R(m-2,n-1) - 0.007\, R(m-2,n-2), \tag{28}$$

with $(m,n) \notin \{(m,n) : 0 \leq m \leq 2,\ 0 \leq n \leq 2\}$.

The borders of the array are assumed to satisfy $(m \geq 5, n \geq 5)$

$$-R(m,0) = -0.713\, R(m-1,0) + 0.105\, R(m-2,0), \tag{29}$$

$$-R(m,1) = -0.926\, R(m,0) - 0.713\, R(m-1,1) + 0.481\, R(m-1,0) + 0.106\, R(m-2,1) - 0.025\, R(m-2,0), \tag{30}$$

$$-R(0,n) = -0.926\, R(0,n-1) + 0.217\, R(0,n-2), \tag{31}$$

$$-R(1,n) = -0.926\, R(1,n-1) + 0.217\, R(1,n-2) - 0.713\, R(0,n) + 0.481\, R(0,n-1) - 0.070\, R(0,n-2). \tag{32}$$

IV. COMPARISON OF SMOOTHING ALGORITHMS

a) "Successive Orthogonalization" and covariance function based on ARMA models. By finding the steady state solution of equation (6), the steady state estimate becomes for the (1,1) ARMA model:

$$\hat{x}(k,s) = 0.04\ \hat{x}(k,s-1) + 0.02\ \hat{x}(k-1,s) + 0.33\ \hat{x}(k-1,s-1) + 0.41\ y(k,s). \quad (33)$$

The mean square value of error is 0.398 and the performance of the filter is given by N = 3.88 db.

Similarly, if one assumes the (2,2) ARMA model, the steady state estimate becomes:

$$\hat{x}(k,s) = 0.049\ \hat{x}(k,s-1) + 0.035\ \hat{x}(k-1,s) + 0.35\ \hat{x}(k-1,s-1) + 0.4\ y(k,s). \quad (34)$$

The mean square value of error is 0.390 and the performance of the filter is given by N = 3.97 db.

Finally, "Successive Orthogonalization" and covariance function based on the (∞,∞) ARMA model the covariance function of equation (20) itself, was considered. The steady state estimate will be:

$$\hat{x}(k,s) = 0.41\ \hat{x}(k,s-1) + 0.34\ \hat{x}(k-1,s) - 0.14\ \hat{x}(k-1,s-1) + 0.33\ y(k,s). \quad (35)$$

The mean square error is 0.32 and the performance of the filter is given by N = 4.83 db.

b) (1,1) ARMA modeling and Optimum Smoothing Algorithm. For the covariance function ((25 to (27)) based on the (1,1) model, equation (13) becomes:

$$\hat{w}(m,n) = y(m,n) - 0.776\ y(m,n-1) - 0.506\ y(m-1,n) + 0.346\ y(m-1,n-1). \quad (36)$$

From this, one can easily determine that

$$\begin{aligned} R_w(m,m+a;\ n,n+b) &= 1.978\ R(a.b) - 0.951\ [R(a,b-1) + R(a,b+1)] \\ &-0.774\ [R(a-1,b) + R(a+1,b)] + 0.393\ [R(a-1,b+1) + R(a+1,b-1)] \\ &+0.346\ [R(a-1,b-1) + R(a+1,\ b+1)]. \end{aligned} \quad (37)$$

Now for $m \geq 1$,

$$T(m,n) = \{(i,j) : i = m,\ 0 \leq j \leq n-1\} \cup \{(i,j) : i = m-1,\ n-1 \leq j\}, \quad (38)$$

while for m = 0,

$$T(0,n) = \{(i,j) : i = 0,\ 0 \leq j \leq n-1\}. \quad (39)$$

Then, for m = 0, equation 16 becomes:

$$R_{w\varepsilon}(0,0;\ n,q) = R_w(0,0;\ n,q) - \sum_{j=0}^{q-1} R_{w\varepsilon}(0,0;\ n,j)\ e^{-1}(0,j)\ R_{w\varepsilon}(0,0;\ q,j)$$

for all $q \leq n$. If $m \geq 1$, p = m, (40)

$$R_{w\varepsilon}(m,m;\ n,q) = R_w(m,m;\ n,q) - \sum_{j=0}^{q-1} R_{w\varepsilon}(m,m;\ n,j)\ e^{-1}(m,j)\ R_{w\varepsilon}(m,m;\ q,j)$$

$$-\sum_{j=n-1}^{D} R_{w\varepsilon}(m,m-1;\ n,j)\ e^{-1}(m-1,j)\ R_{w\varepsilon}(m,m-1;\ q,j), \quad (41)$$

for all $q \leq n$, where D is the width of the picture. If $m \geq 1$, $p = m-1$,

$$R_{w\varepsilon}(m,m-1;\, n,q) = R_{w}(m,m-1;\, n,q) - \sum_{j=n-1}^{q-1} R_{w\varepsilon}(m,m-1;\, n,j)\, e^{-1}(m-1,j)$$
$$R_{w\varepsilon}(m-1,m-1;\, q,j), \tag{42}$$

for all $q \geq n-1$, where $e(m,n) = R_{w\varepsilon}(m,m;\, n,n)$ and the above sum is understood to be zero if the lower limit exceeds the upper limit.

Using this algorithm, the steady state (absolute mean square error variation smaller than 10^{-8}) estimate of w becomes:

$$\hat{w}(m,n) = -0.61\, \varepsilon(m,n-1) - 0.04\, \varepsilon(m,n-2) + 0.28\, \varepsilon(m-1,n-1) - 0.42\, \varepsilon(m-1,n)$$
$$+\ 0.03\, \varepsilon(m-1,n+1). \tag{43}$$

Then

$$\hat{y}(m,n) = 0.16\, y(m,n-1) - 0.04\, y(m,n-2) + 0.09\, y(m-1,n) + 0.06\, y(m-1,n-1)$$
$$+\ 0.03\, y(m-1,n+1) + 0.61\, \hat{y}(m,n-1) + 0.04\, \hat{y}(m,n-2) + 0.42\, \hat{y}(m-1,n)$$
$$-0.28\, \hat{y}(m-1,n-1) - 0.03\, \hat{y}(m-1,n+1). \tag{44}$$

Also,

$$\hat{x}(m,n) = 0.31\, y(m,n) + 0.69\, \hat{y}(m,n). \tag{45}$$

The mean square error is 0.30 and the performance of the filter is given by N = 5.05 db. Note that for this (1,1) ARMA model, five previous estimates and noisy values are needed for the optimum algorithm. These points are shown in figure 5.

V. CONCLUSIONS

For data modeled by the (1,1) ARMA model, which is more realistic than the assumption of a factorable spatial covariance function, the optimum linear recursive estimation procedure requires more than the "previous" three estimates and gives significant improvement over the optimum estimator based on three previous points only. For the particular example considered, the optimum recursive estimator was based on five "previous" points and resulted in a signal-to-noise improvement of 5.05 db as compared to 3.88 db for the estimator based on three "previous"points.

For data with a two-dimensional spatial covariance function that depends on the Euclidean distance between points, reasonable smoothing can be achieved by using an approximate (1,1) ARMA model to fit the data. In order to achieve very near optimal smoothing, however, it appears necessary to employ a rather high order ARMA model which could require considerable off-line calculations to determine the optimum enhancement algorithm. The first point is indicated by the example, where slightly higher improvement was achieved with a (1,1) ARMA fit to the Euclidean covariance function than achieved by the optimal use of the "previous" three estimates based on the actual Euclidean covariance function (4.83 db). The second point is indicated by the improvement achieved based on previous three points under the assumption of a (1,1) ARMA fit (3.88 db), a (2,2) ARMA fit (3.97 db), and the correct covariance function assumption or a (∞,∞) ARMA fit

(4.83 db). It seems reasonable to assume that the optimum smoothing algorithm for the Euclidean covariance function, of the example, would be better (say greater than 6 db improvement) than the algorithm based on the (1,1) ARMA fit.

When dealing with real-world raw data, however, knowledge of the spatial covariance function in closed form cannot be obtained and model identification becomes a difficult problem. Thus, for example, it takes a minimum of 25 estimated values of the covariance function to identify the (2,2) ARMA model. These estimates may appear to satisfy a Euclidean covariance function similar to that of equation (20) and it is possible to estimate the three parameters by a least square fit. Would such a model be more accurate than the (2,2) ARMA model that fit the 25 values exactly? It is not at all clear that the apparent need to employ a higher order ARMA model to fit a Euclidean covariance function similar to that of equation (20) is a real problem when dealing with raw data.

VI. BIBLIOGRAPHY AND REFERENCES

[1] Abend, Harley and Kanal: "Classification of Binary Random Patterns," IEEE Trans. on Information Theory, October 1965.
[2] J.W. Woods: "Two-Dimensional Discrete Markovian Fields," IEEE Trans. on Information Theory, March 1972.
[3] E.Wong: "Two-Dimensional Random Fields and Representation of Images," SIAM J. Appl. Math., Vol. C-21, pp. 756-770, 1968.
[4] E. Wong: "A Likelihood Ratio Formula for Two-Dimensional Random Fields," IEEE Trans. on Information Theory, July 1974.
[5] N.E. Nahi and T. Assefi: "Bayesian Recursive Image Estimation," IEEE Trans. on Computers, Vol. C-21, pp. 734-738, July 1972.
[6] N.E. Nahi and C.A. Franco: "Application of Kalman Filtering to Image Enhancement," Proc. IEEE Conf. Decision and Control, New Orleans, La., pp. 63-65, December 1972.
[7] A Habibi: "Two Dimensional Bayesian Estimate of Images," Proc. of the IEEE, Vol. 60, No. 7, pp. 878-884, July 1972.
[8] S.R. Powell and L.M. Silverman: "Modeling of Two-Dimensional Covariance Functions with Application to Image Restoration," IEEE Trans. on Aut. Control. Vol. AC-19, No. 1, pp. 8-13, February 1974.
[9] M.G. Strintzis: "Comments on Two-Dimensional Bayesian Estimate of Images," under publication in IEEE Proc., 1976.
[10] R. Feick and M.G. Strintzis: "A Recursive Algorithm for Image Enhancement," Proceedings, 12th Annual Allerton Conference on Circuit and System Theory, Monticello, Ill., October 2-4, 1974.
[11] M.G. Strintzis and R. Feick: "ARMA Models of Bivariate Processes," 1976 Pgh. Conf. on Modeling and Simulation.
[12] M.G. Strintzis: "Dynamic Modeling of Cyclic Processes with Application to Two-Dimensional Image Restoration," 1976 Conf. on Information, Sci., and Systems, Johns Hopkins University, pp. 88-90.
[13] H.B. Aasnaes and T. Kailath: "An Innovations Approach to Least-Squares Estimation,Part VIII: Some Applications of Vector Auto-Regressive Moving Average Models," IEEE Trans.Aut.Cont. Vol.AC-18,No.6, pp.601-608, 1973.
[14] A.K. Jain and E. Angel, "Image Restoration, Modelling and Reduction of Dimensionality," IEEE Trans.on Computers,Vol.C-23,No.5,pp.470-476,1974.
[15] M. Strintzis: "Dynamic Representation and Recursive Estimation of Cyclic Processes, with Application to Image Restoration, submitted to IEEE Trans. on Automatic Controls.
[16] M. Kanefsky: "Modeling and Simulation of Two Dimensional Pictures," Pgh. Modeling & Simulation Conf. Vol.6, Part I, pp. 471-473, 1975.

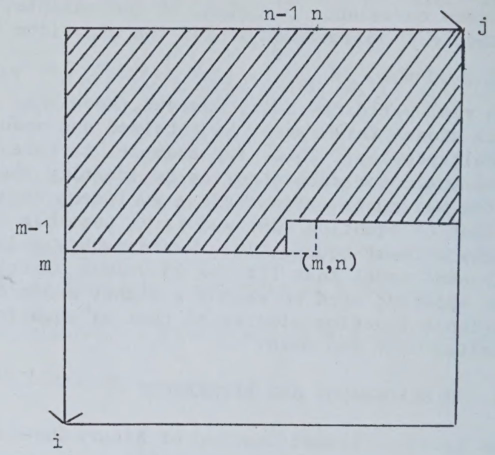
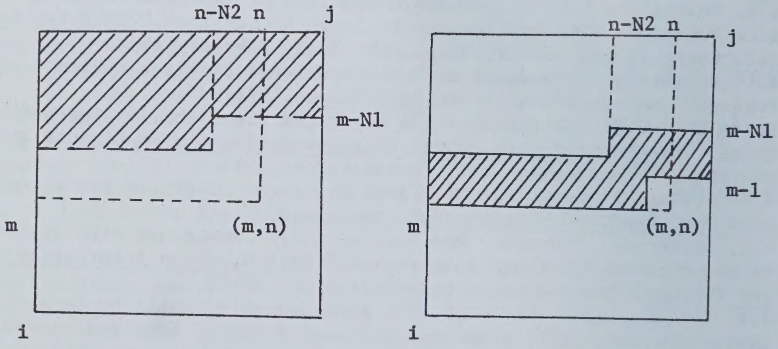

n-1 n
j
m-1
m
(m,n)
i

Figure 1

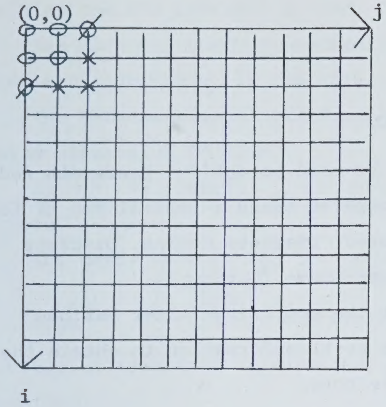
$$S(m,n) = [(i,j):0 \le i \le m-1;0 \le j]\cup[(i,j):i=m;0 \le j \le n-1]$$


n-N2 n
j
m-N1
m
(m,n)
i
n-N2 n
j
m-N1
m-1
m
(m,n)
i

Ψ(m,n)

T(m,n)

Figure 2 Ψ(m,n) and T(m,n)

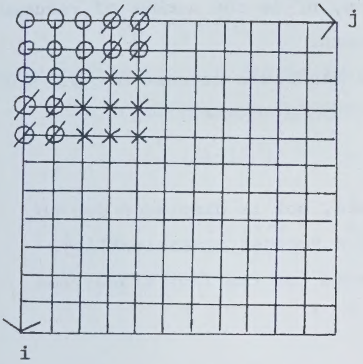


× : indicate the points used to solve " Yule-Walker "equation

○ : indicate the initial values corresponding to Ɏ(m,n)

∅ : indicate additional initial values needed to identify model

Figure 3 R(i,j) for (1,1) ARMA model

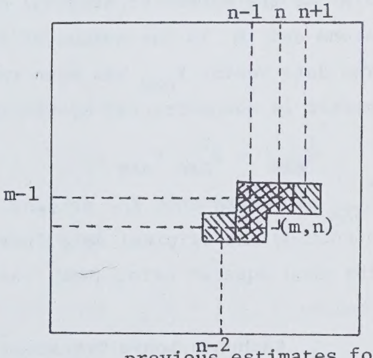


× : same as Figure 3

○ : same as Figure 3

∅ : same as Figure 3

Figure 4 R(i,j) for (2,2) ARMA model



▨ : previous estimates for "successive orthogonalization"

▧ : previous estimates for (1,1)ARMA model

Figure 5

ORTHOGONAL TRANSFORM COMPARISON USING ERTS DATA

GEORGE W. ZOBRIST
The University of Toledo
Toledo, Ohio 43606

ABSTRACT

Discrete orthogonal transforms can be used to achieve bandwidth reduction. Five orthogonal transforms are compared using a typical set of four dimensional ERTS data. The transforms are: Discrete Cosine, Discrete Fourier, Walsh-Hadamard, Haar and Karhunen-Loeve (optimum).

The MSE and clustering are used for the comparison after various levels of dimension reduction and then a re-transformation to obtain the displaced vectors of the original data vectors.

OUTLINE OF RESULTS

The ERTS data used for the orthogonal transform comparison study was obtained from NASA/KSC and consists of a 116 vector subset of a 13260 data set of four dimensional vectors. The data vectors represent four bands of energy, from .5 to 1.1 micrometers, of the shuttle runway area at KSC, Florida.

The reduced dimension data vector Y_{nXM} is obtained from the data vector G_{NXM} by an orthogonal transformation Φ_{nXN},

$$Y_{nXM} = \Phi_{nXN} G_{NXM} ,$$

where N is the number of original dimensions, n is the number of reduced dimensions and M is the number of data vectors.

The data vector Y_{nXM} has been reduced in bandwidth because $n < N$. The data vector is reconstructed by the following inverse operation,

$$\hat{G}_{NXM} = \Phi^{T}_{NXn} Y_{nXM} .$$

$\hat{G}_{NXM}$ now represents the original ERTS data, but is displaced due to reconstructing the original data from data of a reduced dimensionality.

The mean squared error (MSE) was calculated for the five transforms given below:

- Karhunen-Loeve Transform(1)
- Discrete Cosine Transform(2)
- Discrete Fourier Transform(3)
- Walsh-Hadamard Transform(4)
- Haar Transform(5).

The Karhunen-Loeve transform is optimum in the sense that it will minimize the MSE between G and $\hat{G}$ for a fixed number of dimensions.

The Karhunen-Loeve transform is data dependent, while the other transforms are independent of the data.

The Karhunen-Loeve transform is found by finding solutions to the following eigenvalue problem,

$\Sigma_G \, \phi = \lambda\phi$, where Σ_G is the covariance matrix of the original data vectors.

The basis vectors for the discrete Fourier transform are,

$$\phi_{DFT} = e^{-j\frac{2\pi nk}{N}} , \; n = 0, 1, \ldots, N-1; \quad k = 0, 1, \ldots, N-1.$$

The Fourier transform is asymptotically equivalent to the Karhunen-Loeve transform.

The basis vectors for the discrete cosine transform are,

$$\phi_{DCT} = \sqrt{1/2} \qquad n = 0, 1, \ldots, N-1$$

$$\phi_{DCT} = \cos[(2n+1)k\pi/2N], \quad \begin{array}{l} n = 0, 1, \ldots, N-1 \\ k = 1, \ldots, N-1 \end{array}$$

The discrete cosine transform is a Fourier series representation of periodically extended data.

The basis vectors for the Walsh-Hadamard transform are,

$$\phi_W = (-1)^{\sum_{i=0}^{N-1} u_i x_i}$$

, where u_i and x_i are binary bits representing the row and column indices.

The basis vectors for the Haar transform are,

$$\phi_H = \begin{cases} 2^{(k-1)/2} & ; \; \frac{n-1}{2^{k-1}} < \tau \leq \frac{n-1/2}{2^{k-1}} \\ -2^{(k-1)/2} & ; \; \frac{n-1/2}{2^{k-1}} < \tau \leq \frac{n}{2^{k-1}} \\ 0 & ; \; 0 < \tau \leq \frac{n-1}{2^{k-1}} \text{ and } \frac{n}{2^{k-1}} < \tau \leq 1 \end{cases}$$

where τ is in the interval [0,1] and is subdivided into N equal increments. The Haar transform tends to overcome error propagation, since it is both locally and globally sensitive.

Results of the comparison of the five transforms are presented as graphs of the mean squared error for each data vector.

REFERENCES

1. Fukunaga, K. and Koontz, W. L. G., "Application of the Karhunen-Loeve Expansion to Feature Selection and Ordering," Trans. IEEE, C-19, 1970, pp. 311-318.
2. Ahmed, N., Natarajan, T., and Rao, K. R., "Discrete Cosine Transform," Trans. IEEE Comp., January 1974, pp. 90-93.
3. Pearl, J., "On Coding and Filtering Stationary Signals by Discrete Fourier Transform," Trans. IEEE, IT-19, March 1973, pp. 229-232.
4. Pratt, W. K., Kane, J., and Andrews, H. C., "Hadamard Transform Image Coding," Proc. IEEE, vol. 57, no. 1, January 1969.
5. Andrews, H. C., "A Generalized Technique for Spectral Analysis," Trans. IEEE, C-19, January 1970, pp. 16-25.

ACKNOWLEDGEMENTS

This is to acknowledge the support of Mr. Bristol Harris, NASA/KSC, who prepared the ERTS data tapes, and Mr. John Jamison, NASA/KSC, who did the computer programming.

NOVEL USES OF QUADRATIC SURFACES FOR MEDICAL DIAGNOSIS*

N. DE CLARIS, O. IJAOLA, R. W. NEWCOMB
University of Maryland, College Park, Maryland 20742

ABSTRACT

There are situations in medical diagnostics where conventional clustering and pattern recognition techniques are impractical or all together they fail to apply. For these cases the concept of "minimal ellipsoids" is introduced together with an approach for its utlization in disease classification and illness identification in a given patient.

I. Introduction

The use of information processing machines in medicine to process patient data, to recognize patterns and to screen and diagnose disease(1, 2) is increasing. Generally it is based on two broad approaches(3):
a) pattern recognition using probabilistic models and statistical analysis,
b) automatic classification schemes emphasizing accuracy. However, there are situations in clinical diagnosis where no collection of statistical data can provide the diagnostic precision needed to correctly identify the illness for therapeutic purposes(4). The reasons why in these situations statistical techniques are unsatisfactory can be well understood(5, 6). A deterministic approach is outline here for separating clinical and paraclinical data of patients into classes appropriate for diagnostic and therapeutic purposes. The heart of the approach is a novel use of quadratic surfaces(7, 8) for "fine classification" of diseases and their subsequent use to identify "specific illness" of a given "host".

We show that the approach is: (a) well formulateable mathematically (in terms of well posed problems for which solutions do exist) and (b) analytically numerically and computationally tractable.

The following is a problem which arises in medical (differential) diagnosis. From confirmed clinical past cases it is known that certain physiological measurements (symptom, observations, laboratory tests, etc.) receive particular sets of values for an identified "disease A" (abnormal state of the physiological system) and a different set of values for another distinct "disease B". Given that these data are available, how does a physician go about deciding whether a given patient whose phys-

---

[*] Supported in part by U.S. Public Health Service Grant, No. RRO-7042-11.

iological measurements he just obtained is "host" to "disease A" or to "disease B" or he is facing a new case that can not be decided upon by the present measurements and the past data?

The motivation and fundamental theoretical considerations which led to our mathematical formulation of this problem are best visualized geometrically in two dimensions. Thus our discussion is initially made with reference to the x-y plane, however it should perhaps be amphasized from the beginning that the practical cases of interest (for medical diagnosis) involve situations whose representations requires much higher dimensionality, of the order 20 to 200.

Thus, given N cases of known "disease A" they may be viewed as N points (x_i^α, y_i^α) on the x-y plane forming the set A while M cases of known "disease B" cases may be viewed as points (x_j^b, y_j^b) of a second set B. That is

$$(x_i^\alpha, y_i^\alpha) \varepsilon A \qquad i = 1, 2, \ldots N \tag{1.a}$$

and

$$(x_j^b, y_j^b) \varepsilon B \qquad j = 1, 2, \ldots M \tag{1.b}$$

under the assumption that sets A and B cluster near points

$$(x_o^\alpha, y_o^\alpha) \text{ and } (x_o^b, y_o^b)$$

respectfully, for convenience.

A simple and frequently mentioned technique is to enclose sets A and B each with a circle C_A and C_B defined by the radii r_A and r_B respectively. Then to decide whether a new case - a point (x^t, y^t) - is: 1) inside C_A, 2) inside C_B, or 3) outside C_A and of C_B.

Mathematically this amounts to the use of the quatratic form:

$$(x^t - x_o)^2 + (y^t - y_o)^2 \begin{cases} > r_A^2 & \text{outside } C_A \\ \leq r_A^2 & \text{inside } C_A \end{cases} \tag{2.a}$$

where

$$r_A^2 = \max_{\forall_i} \left[(x_i^\alpha - x_o)^2 + (y_i^\alpha - y_o)^2 \right] \tag{2.b}$$

to decide wheter the given point (x^t, y^t) is inside or outside the circle of radius r_A. This simple approach is sueful (affording much computational convenience) when the circles C_A and C_B do not intersect on the plane, Figure 1. Clearly the approach is of limited applicability if the two circles happen to intersect while the given data do not; Figure 2. While it is possible to pre-stretch the axis through a linear transformation, and then define the circles C_A and C_B this (eigenvalue scheme) does not always work for the same basic reason.

A much better appraoch is to enclose, if possible, the given data points (sets A and B) with non-intersecting ellipses, E_A and E_B, Figure 3, and then use these quadratic forms to test the point (x^t, y^t). The question then in two dimensions is: given a set of points $N \geq 3$, is it possible to find an ellipse E, containing all points and enclosing minimum area (points on E are considered as contained by the ellipse)?

It is not hard to show that this is equivalent to using the quadratic form (in the z-w plane, a transformation of the x-y plane):

$$\left(\frac{z}{d_1}\right)^2 + \left(\frac{w}{d_2}\right)^2 \begin{cases} > 1 & \text{outside E} \\ \leq 1 & \text{inside E} \end{cases} \qquad (3)$$

where d_1 and d_2 are the minimum values possible and defined by three suitable points of the given set.

In the next two sections we outline the much harder problems of:

1. establishing that this basic approach can be extended to n-dimensions, and
2. formulating practical algorithms (for computer implementations) for finding the "minimal ellipsoid" from given sets od data and then utilizing it for diagnosing a particular case under question.

There are still several other aspects that must be considered before the effectiveness of the approach can be taken full advantage of for general medical diagnosis purposes, such as updating the data, modification by learning, etc. However these problems, under active considerations are not taken up in this paper.

II. Mathematical Formulation

The data to be classified are considered, in the usual fashion$^{(7,8)}$ as n-vectors in the real domain R^n. If the scalar product of two vectors x, y is written as $\langle x, y\rangle$ and K is an nxn positive definite (real symmetric) matrix, the equation

$$\langle x, Kx\rangle \leq 1 \qquad x \in R^n \qquad (4)$$

defines an n-dimensional ellipsoid. The main idea is, given a set of data vectors:

$$\alpha^i \in R^n \qquad i = 1, 2, 3, \ldots, N \geq n \qquad (5)$$

there exists a "minimal ellipsoid" which contains the data and which has a maximum number of data points on its boundary. In fact it can be shown$^{(9)}$ that the minimal ellipsoid is completely determined by n+1 data points which must lie on its surface.

The situation is somewhat simplified by finding an orthongonal matrix

A which diagonalizes K so that equation (5) becomes:

$$\langle y, D^{-1} y \rangle \leq 1 \tag{6.a}$$

where

$$AKA^{tr} = D^{-1} \tag{6.b}$$

and

$$D = \text{diag.}\,[d_i] \qquad i = 1, 2, \ldots, n \tag{6.c}$$

That is we need to find suitable orthogonal coordinates in R^n (with reference to the data) to identify d_i. This is accomplished by an original axis translation and successive axis rotation (about the origin) ensuring that all given data vectors fall within the volume defined by the ellipsoid.

Thus given a set of α^i a new set is generated

$$\phi^i = \alpha^i - c_o \tag{7.a}$$

and it is reordered from

$$||\phi^i|| = n_i \tag{7.b}$$

to satisfy $n_1 \geq n_i$ $i = 1, 2, 3, \ldots, N$, where $||\phi^i||$ denotes the usual norm. C_o is also chosen as to minimize the max n_i for all i. Next rotation of the axis takes place to align the first axis along ϕ_1. Then through suitable rotation and the Gram-Schmidt procedure a square matrix R is found such that

$$RR^{tr} = 1 \qquad \text{and} \tag{8.a}$$

$$R\phi^1 = y_1 = \begin{bmatrix} n_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{8.b}$$

we choose $d_1 = n_1^2$ and proceed in the same mannor to find the other coordinates. However, we must be careful to rotate all remaining components (after we extract the vector components along the selected axis) about the just selected axis and test using the sub ellipsoid (quadratic surface) for the coordinates selected. This is accomplished through the equation:

$$d_2 = \max_i \left\{ \frac{(n_i^{(2)})^2}{1 - [(\phi_{i1}^{(2)})^2 / d_i]} \right\} \tag{9}$$

This process repeats itself by successive rotations and gives explicit expression for the overall transformation and the matrix $D^{(9)}$. The notation does get a bit complicated but the problem is straightforward.

III. Algorithmic Aspects

In this section an algorithm for constructing "minimal ellipsoids" from a given set of data is outlined in much simplified from in order to illuminate the basic features of the approach. Some other desirable features are also present, such as provisions for identifying possible multi-ellipsoidal surfaces and/or learning capability (perception like), but they are not discussed here.

The first part of the algorithm sets up suitable notation, Figure 4, where:

α^i represent confirmed patients as vectors with reference to measurement axis x_j, $j = 1, 2, \ldots (N+1)$ and $i = 1, 2, \ldots$ $M > N+1$

y_k represent unit vectors coordinate axis defining with reference to the minimal ellipsoid, $k = 1, 1, \ldots N$

β_k^i represent the confirmed patient vector data projected on the y_k defined axis $i = 1, 2, \ldots N+1$ and $k = 1, 2, \ldots N$.

Remark: The equation for the interior of the minimal ellipsoid is selected so as to satisfy:

$$\sum_{k=1}^{N} \left(\frac{\beta_k^i}{d_k} \right)^2 \leq 1 \qquad \text{for all } j \tag{10}$$

with d_k having the smallest possible values. The procedure is to select one y_k, test all β_k^i $(i = 1, 2, \ldots N+1)$ and choose d_k to satisfy the above inequality, then go back and select another y_{k+1} (orthogonal to y_k). This process is repeated until all y_k have been found.

The first axis (the major axis) of the ellipsoid is found by constructing the difference vectors

$$v^r = \alpha^i - \alpha^j \qquad r = 1, 2, \ldots, \frac{N(N+1)}{1}$$

The vector v^R is given by:

$$|| v^R ||^2 = (2d_1)^2 \geq || v^r || \text{ for all } r$$

establishes $y_1 = \frac{1}{2d_1} v^R$ and of course d_1.

The second part of the algorithm, Figure 5, has two branches. One branch performs the expansion

$$
\begin{aligned}
C^j &= \beta_1^j y_1 + z_1^j \\
z_1^j &= \beta_{2y_2}^j + z_2^j \\
&\cdots\cdots\cdots\cdots \\
z_{N+1}^j &= \beta_N^j y_N + z_N^j
\end{aligned}
\tag{11}
$$

where $\langle y_k \cdot z_k^j \rangle = 0$. The other branches main subroutine is the SHRINK TEST where each z_{k-1}^j is tested via the inequality:

$$
\sum_{i=1}^{k-1} \left(\frac{\beta_i^j}{d_i} \right)^2 + \left(\frac{\| z_{k-1}^j \|}{(1 - \rho\Delta) d_{k-1}} \right)^2 \begin{cases} > 1 & \text{outside} \\ \leq 1 & \text{inside} \end{cases}
\tag{12}
$$

and ρ is increased one step at a time $\rho = 1, 2, \ldots, R$.

The $\hat{z}_{k-1}^j$ which ensures the maximum value of ρ_{mx} and all data points inside the ellipsoid determines y_k and d_k. That is:

$$
y_k = \frac{1}{\| \hat{z}_{k-1}^j \|} \hat{z}_{k-1}^j
\tag{13}
$$

and

$$
d_k = (1 - \rho_{mx} \Delta) \, d_{k-1}
\tag{14}
$$

Remark: $\Delta = \frac{d_1}{R}$ is a predetermined unit of axis "shrinkage", and R the total number of "shrinkage steps".

IV. Concluding Remarks

Various parts and slightly different versions of the above algorithm have been successfully tested computationally. The computer implementation and actual medical use of this approach involves additional considerations (computational efficiency, storage, accuracy, etc.) beyond the scope of this paper.

V. References

1. E. R. Gabrieli, "The Use of Data Mechanization and Computers in Clinical Medicine", N. Y., Acad. of Sci., 1969.
2. E. A. Patrick, et. al., "Review of Pattern Recognition in Medical Diagnosis and Consulting Relative to a New System Model", IEEE Syst., Man., Cybern., January 1974.
3. C. A. Kulikowski, "Pattern Recognition Approach to Medical Diagnosis", IEEE Trans. Syst., Sci., Cybern., July 1970.
4. A. R. Feinstein, "Clinical Judgement", The Williams and Wilkins Co., Baltimore, 1969.

5. L. Kanal, "Patterns in Pattern Recognition: 1968-1974", IEEE Trans. on Information Theory, November 1974.

6. K. Fukunaga, "Introduction to Statistical Pattern Recognition", Academic Press, New York, 1972.

7. N. J. Nilson, "Learning Machines", McGraw-Hill.

8. R. O. Duda and P. E. Hart, "Pattern Classification and Scene Analysis", Wiley, New York, 1973.

9. O. Ijaola, Ph. D. Thesis, Graduate School, University of Maryland, (under preparation).

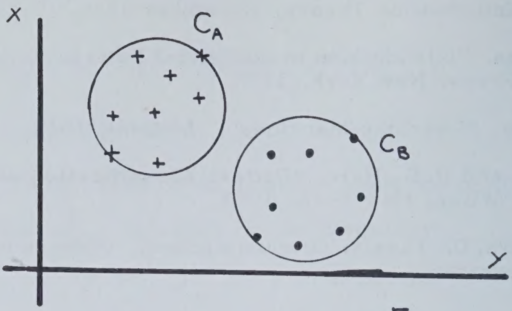


Figure 1

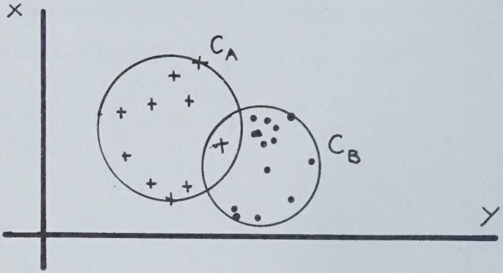


Figure 2

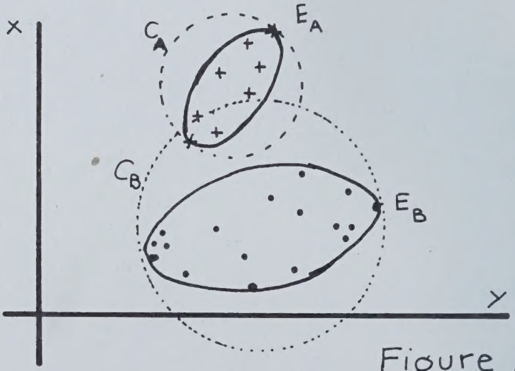

X
C_A
E_A
C_B
E_B
Y

Figure 3

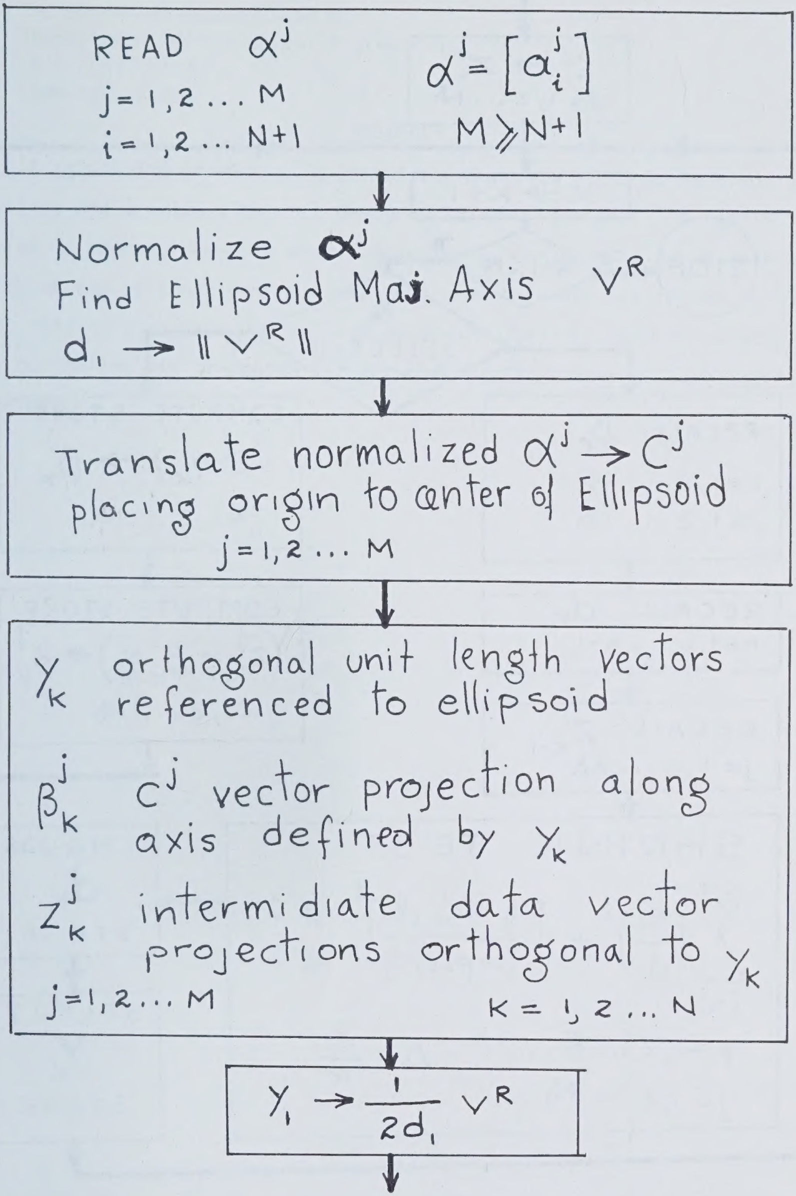

READ α^j
$j = 1, 2 \ldots M$
$i = 1, 2 \ldots N+1$
$\alpha^j = [a_i^j]$
$M \geqslant N+1$
Normalize α^j
Find Ellipsoid Maj. Axis V^R
$d_1 \rightarrow \| V^R \|$
Translate normalized $\alpha^j \rightarrow C^j$
placing origin to center of Ellipsoid
$j = 1, 2 \ldots M$
Y_k orthogonal unit length vectors referenced to ellipsoid
β_k^j C^j vector projection along axis defined by Y_k
Z_k^j intermidiate data vector projections orthogonal to Y_k
$j = 1, 2 \ldots M$
$K = 1, 2 \ldots N$
$Y_1 \rightarrow \frac{1}{2d_1} V^R$

Figure 4

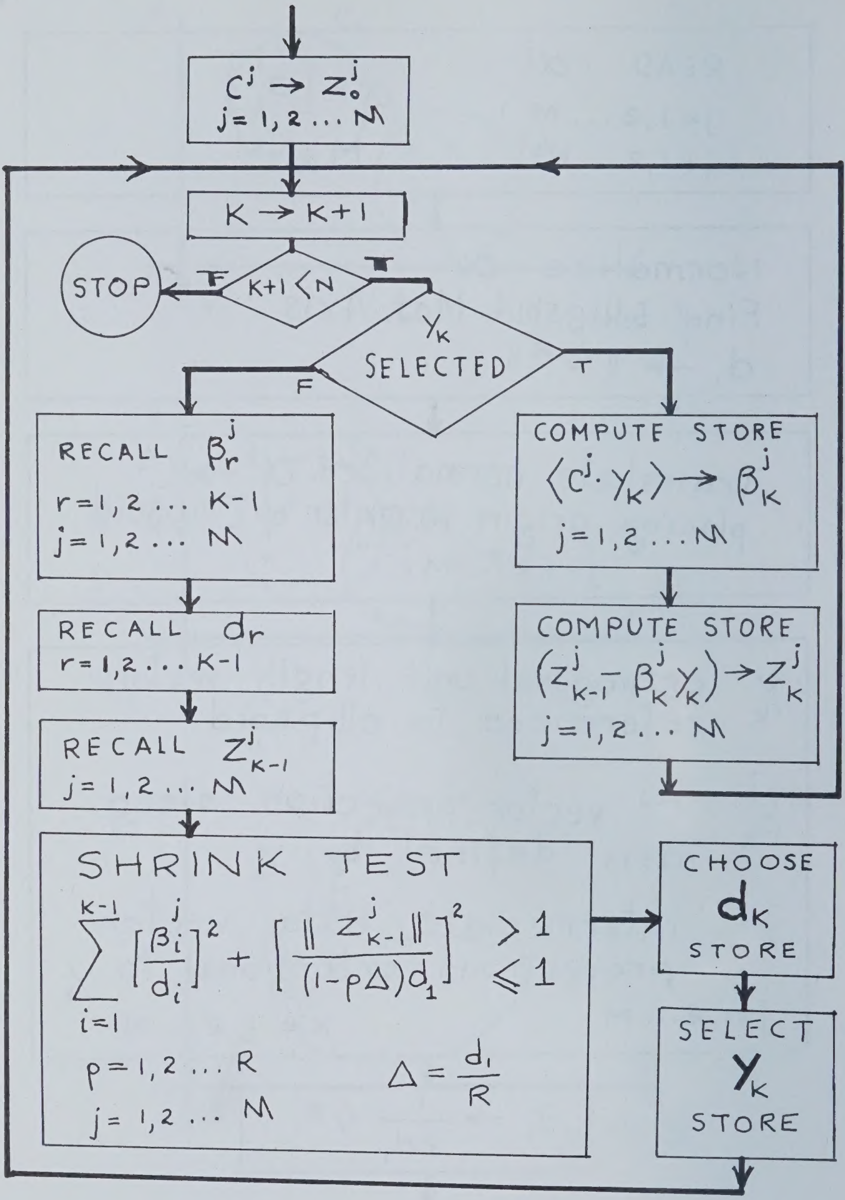

$C^j \rightarrow Z_0^j$
$j = 1, 2 \ldots M$
$K \rightarrow K+1$
STOP
F
$K+1 < N$
T
Y_K SELECTED
F
T
RECALL β_r^j
$r = 1, 2 \ldots K-1$
$j = 1, 2 \ldots M$
COMPUTE STORE
$\langle C^j \cdot Y_K \rangle \rightarrow \beta_K^j$
$j = 1, 2 \ldots M$
RECALL d_r
$r = 1, 2 \ldots K-1$
COMPUTE STORE
$(Z_{K-1}^j - \beta_K^j Y_K) \rightarrow Z_K^j$
$j = 1, 2 \ldots M$
RECALL Z_{K-1}^j
$j = 1, 2 \ldots M$
SHRINK TEST
$\sum_{i=1}^{K-1} \left[\frac{\beta_i^j}{d_i}\right]^2 + \left[\frac{\| Z_{K-1}^j \|}{(1-\rho\Delta)d_1}\right]^2 \begin{matrix} > 1 \\ \leq 1 \end{matrix}$
$\rho = 1, 2 \ldots R$
$\Delta = \frac{d_1}{R}$
$j = 1, 2 \ldots M$
CHOOSE d_K STORE
SELECT Y_K STORE

Figure 5

A SEQUENTIAL PROCEDURE FOR COSTLY DATA CLUSTERS

P. PAPANTONI-KAZAKOS and R. LI
Department of Electrical Engineering
Rice University
Houston, Texas

ABSTRACT

A sequential procedure is formulated for data received in fixed size clusters and a collection cost assigned to each cluster. The clusters are finite in number, they are used sequentially, and a decision is made as to which of two alternatives is true. The procedure and its optimality vary with varying statistical knowledge about the system. Conditions guaranteeing optimality are presented for three different stages of such statistical knowledge.

This work was supported by NSF Grant ENG 75-10516

A SEQUENTIAL PROCEDURE FOR COSTLY DATA CLUSTERS

P. Papantoni-Kazakos and R. Li

INTRODUCTION

Let hypothesis H_1 is tested against hypothesis H_o. The group x_1 that consists of k_1 data is first collected and used for the hypothesis testing. At this stage either a decision is made and no more data are collected or the decision is postponed and the second data cluster x_2 is collected. In general, group x_i (consisting of k_i data) is collected if and only if no decision could be made before the i^{th} stage. Also, at each stage an effort is made for a decision against or for gypothesis H_1.

The data clusters are, in general, n where n is a finite number and the sizes k_i are well known. Furthermore, a collection cost c_i is assigned to the data cluster x_i where the constants c_i are positive and known for each $1 \leq i \leq n$. Finally, penalties for wrong decision are also given. In particular, c_{k_j} will be the positive penalty for deciding H_k, while H_j is true where $0 \leq k, j \leq 1$.

With the above data remaining unchanged, the sequential scheme will take different forms for the different stages of statistical knowledge about the model.

Denote by X_i the vector of the data collected at stage i. That is

$$X_i^T = [x_1^T, x_2^T, \dots, x_i^T]$$

where the superscript T means transpose.

The procedure we present here consists of three threshold parameters. A, B, K such that $A < B$. The threshold K is needed to terminate the procedure for finite number n of data clusters and the scheme works as follows:

i) The data vector X_i is collected if and only if no decision could be made before the i^{th} stage.

ii) The function $T(X_i)$ is calculated where $T(\cdot)$ is a real, scalar function which is part of the procedure and will be designed optimally.

iii) For $i \leq n - 1$ if the data are such that $A < T(X_i) < B$, no decision is made and the $(i + 1)^{th}$ data group is collected.

iv) For $i \leq n - 1$ if $T(X_i) = A$ or $T(X_i) = B$, it is decided that H_o is true with probability P_A or that H_1 is true with probability P_B correspondingly and the procedure ends.

v) For $i \leq n - 1$, if $(TX_i) < A$, H_o is decided and the procedure ends.

vi) For $i \leq n - 1$, if $T(X_i) > B$, H_1 is decided and the procedure ends.

vii) The threshold K is used if and only if no decision could be made before the n^{th} stage. Then, if $T(X_n) > K$, H_1 is decided; if $T(X_n) = K$, H_1 is decided with probability P_K; and if $T(X_n) < K$, H_o is decided.

As a conclusion, the procedure is a set $\{T(\cdot), A, B, K, P_A, P_B, P_K\}$. The constants, as well as the function $T(\cdot)$ are designed optimally where the optimality criterion varies with the statistical knowledge about the system.

Three cases of such knowledge are considered and the sequential scheme is designed accordingly.

1. Channel statistics $P_1(X_i)$, $P_o(X_i)$ under both hypotheses H_1, H_o well known for every i. A priori probabilities q_1, q_o of the two hypotheses also known. Then, the optimally criterion becomes Bayesian and unique optimal procedure exists under certain conditions.

It is found that the optimal function $T^o(\cdot)$ and the threshold K^o can be found unconditionally and they are given by:

$$T^o(X_i) = \frac{dP_1(X_i)}{dP_o(X_i)} \; ; \; K^o = \frac{q_o}{q_1} \frac{c_{10}}{c_{01}}$$

On the contrary, unique optimal A^o, B^o such that $A^o < B^o$ exist only under conditions. Such efficient conditions are expressed by corollary 1 for two data clusters and smooth but otherwise general channel statistics.

Corollary 1

Let $T^o(X_1)$, $T^o(X_2)$ be jointly absolutely continuous random variables and let $f_{2,1/H_i}(y,x)$, $f_{1/H_i}(x)$, $f_{2/1,H_i}(y/x)$ be the corresponding joint, marginal and conditional density functions. In addition, let the marginal densities $f_{1/H_i}(x)$, $f_{1/H_o}(x)$ be differentiable, the conditional density $f_{2/1,Hi}(y/x)$ be such that

$$\frac{\partial f_{2/1,Hi}(y/x)}{\partial x} = - D \frac{\partial f_{2/1,Hi}(y/x)}{\partial y} \tag{1}$$

where D is a positive constant,

$$\frac{\partial}{\partial x} \int f_{2/1,H_i}(y/x)\, dy = \int \frac{\partial}{\partial x} f_{2/1,H_i}(y/x)\, dy \tag{2}$$

and

$$\inf_x \left[\frac{f'_{1/H_1}(x)}{f_{1/H_1}(x)} - \frac{f'_{1/H_0}(x)}{f_{1/H_0}(x)} \right] > 0 \tag{3}$$

where $f'(x)$ is the first order derivative with respect to x. Then, if

$$\min (c_{01}, c_{10}) > C_2 > \frac{D}{\inf\limits_x \left[\frac{f'_{1/H_1}(x)}{f_{1/H_1}(x)} - \frac{f'_{1/H_0}(x)}{f_{1/H_0}(x)} \right]}$$

$$\max \left(c_{01} \sup_x f_{2/1,H_1}(K^\circ/x),\ c_{10} \sup_x f_{2/1,H_0}(K^\circ/x) \right) \tag{4}$$

and there is an x value x° such that

$$c_2 < \left[\min \left(1, \frac{f_{1/H_1}(x^\circ)}{K^\circ\ f_{1/H_0}(x^\circ)} \right) - \int_{K^\circ}^{\infty} f_{2/1,H_0}(y/x^\circ) dy - \frac{f_{1/H_1}(x^\circ)}{K^\circ\ f_{1/H_0}(x^\circ)} \cdot \int_{-\infty}^{K^\circ} f_{2/1,H_1}(y/x^\circ)\, dy \right] \cdot \left(\frac{1}{c_{10}} + \frac{1}{c_{01}\ K} \frac{f_{1/H_1}(x^\circ)}{f_{1/H_0}(x^\circ)} \right)^{-1} \tag{5}$$

there are unique optimal thresholds A°, B° such that $A^\circ < B^\circ$.

These optimal thresholds are given then by the unique zeros of the following two functions correspondingly.

$$g(x) = c_2 \left[\frac{1}{c_{10}} + \frac{1}{c_{10}\frac{q_0}{q_1}} \frac{f_{1/H_1}(x)}{f_{1/H_0}(x)} \right] - 1 + \int_{K^\circ}^{\infty} f_{2/1,H_0}(y/x)\, dy + \\ + \frac{c_{01}}{c_{10}\frac{q_0}{q_1}} \frac{f_{1/H_1}(x)}{f_{1/H_0}(x)} \int_{-\infty}^{K^\circ} f_{2/1,H_1}(y/x)\, dy \tag{6}$$

$$h(x) = -c_2 \left[\frac{1}{c_{10}} + \frac{1}{c_{10}\frac{q_0}{q_1}} \; \frac{f_{1/H_1}(x)}{f_{1/H_0}(x)} \right] + \frac{c_{01}}{c_{10}\frac{q_0}{q_1}} \; \frac{f_{1/H_1}(x)}{f_{1/H_0}(x)} -$$

$$- \frac{c_{01}}{c_{10}\frac{q_0}{q_1}} \; \frac{f_{1/H_1}(x)}{f_{1/H_0}(x)} \int_{-\infty}^{K^\circ} f_{2/1,H_1}(y/x)\, dy - \int_{K^\circ}^{\infty} f_{2/1,H_0}(y/x)\, dy \tag{7}$$

2. Channel statistics $P_1(X_1)$, $P_0(X_1)$ well known, but no a priori probabilities available. The optimality criterion is then the minimization of the average cost conditioned on hypothesis H_1 under the restriction that the average cost conditioned on H_0 is equal to a constant. Lagrange multipliers methods are applied and it is found again that the optimal $T^0(.)$, K^0 are

$$T^0(X_1) = \frac{dP_1(X_1)}{dP_0(X_1)} \; ; \; K^0 = (-\lambda) \; \frac{c_{10}}{c_{01}}$$

where λ is the Lagrange multiplier while A^0, B^0 exist uniquely only under certain restrictions. Again such sufficient conditions are found for general smooth channel statistics, and are expressed by corollary 2.

Corollary 2

For $f_{2,1/H_1}(y,x)$, $f_{1/H_1}(x)$, $f_{2/1,H_1}(y/x)$ expressing the same densities first presented by corollary 1 for $\lambda < 0$ and for conditions (1-4) satisfied, then if there is a $x^\circ(\lambda) = x^\circ$ such that (5) is also true, there are unique optimal thresholds A°, B° given by the zeros of the functions (8) and (9) correspondingly and such that $A^\circ < B^\circ$, where:

$$g(x) = c_2 \left[\frac{1}{c_{10}} + \frac{1}{c_{10}(-\lambda)} \; \frac{f_{1/H_1}(x)}{f_{1/H_0}(x)} \right] - 1 + \int_{K^\circ}^{\infty} f_{2/1,H_0}(y/x)\, dy +$$

$$+ \frac{c_{01}}{(-\lambda)\, c_{10}} \; \frac{f_{1/H_1}(x)}{f_{1/H_0}(x)} \int_{-\infty}^{K^\circ} f_{2/1,H_1}(y/x)\, dy \tag{8}$$

$$h(x) = -c_2 \left[\frac{1}{c_{10}} + \frac{1}{c_{10}(-\lambda)} \; \frac{f_1/H_1(x)}{f_1/H_0(x)} \right] + \frac{c_{01}}{c_{10}(-\lambda)} \; \frac{f_1/H_1(x)}{f_1/H_0(x)} -$$

$$- \frac{c_{01}}{c_{10}(-\lambda)} \frac{f_{1/H_1}(x)}{f_{1/H_0}(x)} \int_{-\infty}^{K^\circ} f_{2/1,H_1}(y/x)\, dy - \int_{K^\circ}^{\infty} f_{2/1,H_0}(y/x)\, dy \tag{9}$$

3. The channel statistics form a nonparametric family here. Tests are sought whose characteristics remain unchanged inside the whole nonparametric family. The characteristics $\{T(\cdot),A,B,K,P_A,P_B,P_K\}$ of the procedure are determined after the "nonparametric" sequential procedure is defined precisely. As a measure of evaluation of the procedure in the actual presence of certain parametric channel statistics, an extended version of the asymptotic relative efficiency is defined and certain "nonparametric" tests are analyzed.

Let us define by R_0 the average risk conditioned under H_0. Also, let

$$R_0\ell = c_{10}\ \Pr\{T(X_1) > B/H_0\} + c_{10}\ P_B\ \Pr\{T(X_1) = B/H_0\} \tag{10}$$

be the average risk paid by the system when under H_0, H_1 is decided at the collection of the first data cluster. $R_0\ell$ bounds R_0 from below. Finally, let

$$R_{01} = c_2 + c_{10}\ P_K\ \Pr\{T(X_2) = K/H_0\} + c_{10}\ \Pr\{T(X_2) > K/H_0\} \tag{11}$$

be the average risk paid by the system when collection cost is assigned only to the second data cluster and at such collection H_1 is decided under condition H_0. The sum $R_0\ell + R_{01}$ bounds R_0 from above.

Definition 1

A sequencial two-cluster test $\{T(\cdot),A,B,K,P_A,P_B,P_K\}$ is called non-parametic inside a nonparametic distribution family if and only if the risks $R_0, R_{0\ell}, R_{01}$ remain unchanged inside this family.

It must be pointed out here that the above sequential nonparametric test results in different R_1 values (where R_1 is the average risk conditioned on H_1) for different distributions in the nonparametric family. R_1 can be used as a performance measure of the test at each of the above distributions. Similarly to the nonsequential nonparametric test, an asymptotic relative efficiency (A.R.E.) can be assigned here also. This efficiency will be based on the risk R_1 and it is defined as follows.

Definition 2

The asymptotic relative efficiency of a certain two-cluster nonparametric test at a given distribution F_0 and with respect to a given parametric sequential two-cluster test is defined as the ratio of the common for the two clusters number of data k_n required by the nonparametric test and the distribution F_0 to achieve given R_0, R_1 risks with equally sized clusters and the two hypothesis H_0, H_1 approaching each other, over the number of data k_p required by the considered parametric test and the same distribution F_0 to achieve the same R_0 and R_1 with equally sized clusters and $H_0 \rightarrow H_1$ also.

Once the conditional risks R_0, R_1 are specified, the parametric test is determined through minimization of R_1 conditioned on the fixed value of R_0. The nonparametric test is determined through given $R_0, R_0\ell, R_{01}$ values. The assigned R_1 value is used for the specification of the ratio $\frac{k_p}{k_n}$. Also, since $H_0 \rightarrow H_1$ it is necessary that both k_p and k_n become very large for the specified R_0, R_1 values to be maintained. The same $T(\cdot)$ test functions that are nonsequentially nonparametric are sequentially nonparametric too. The parametric sequential test that is used as a basis for comparison at the evaluation of nonparametric sequential tests is the unknown priors two-cluster one presented as case 2. The distribution to at which the sequential nonparametric tests is evaluated is the Gousian one.

There is no space here for the presentation of the complete evaluation analysis. It will be only stated that conditions similar to the ones in corollary 1 guarantee the existence of unique optimal thresholds A°, B° such that $A^\circ < B^\circ$. The calculation of the A.R.E. as defined here follows a procedure that is similar for the sign and Wilcoxon nonparametric tests. The ratio $\frac{k_p}{k_n}$ can be found only numerically in all these cases due to the fact that this is the only way to obtain the optimal thresholds A_1°, B_1° of the parametric test.

References

1. Wald, A. (1947), Sequential Analysis, Wiley, New York.

2. Wald, A. and J. Wolfowitz (1948) "Optinum Character of the Sequential Probability Ratio Test," Ann. Math. Statist. 19, 326-339.

3. M. Sobel and A. Wald (1949), "A Sequential Decision Procedure for choosing one of Three Hypothesis Concerning the Unknown Mean of a Normal Distribution" Ann. Math. Statist. 20, 502-522.

4. P. Armitage, (1950) :Sequential Analysis with more than two alternative hypothesis and its relation to discriminant function analysis," J. Roy. Statist. Soc. Ser. B 12, 137-144.

5. D.R. Cox, (1952) "Sequential Tests for Composite Hypothesis," Proc. Cambridge Phil. Soc. 48, 290-299.

6. A. Droretzky, J. Kiefer and J. Wolfomitz, (1953) "Sequential Decisions Problems for Processes with Continous Time Parameter, Testing Hypothesis," Ann. Math. Statist. 24, 254-264.

7. T.W. Anderson, (1960) "A Modification of the Sequential Probability Ratio Test to Reduce the Sample Size," Ann. Math. Statist. 31, 165-197.

8. F.C. Reed, (1960) "A Sequential Multi-Decision Procedure," Proc. Symp. on Decision Theory and Appl. Electron. Equipment Develop., USAF Develop. Center, Rome, New York.

9. R. Bellman, R. Kalaba, and D. Middleton (1961)"Dynamic Programming, Sequential Estimation and Sequential Detection Processes," Proc. Nat. Acad. Sci. 47, 338-341.

10. D.R. Cox (1963) "Large Sample Sequential Tests for Composite Hypothesis," Sankhya 25, 5-12.

11. E.B. Dynkin, (1967) "The Optimum Choice of the Instant for Stopping a Markov Process," Soviet Math. Dokl. 4, No. 3, 627-629.

12. R.M. Phataford, (1963) "Large Sample Sequential Analysis of Markovian Observations," J. India Statist. Assoc. 1, No. 3, 152-160.

13. F. Wilcoxon, L.J. Rhodes and R.A. Bradley, (1963) "Two Sequential Two-Sample Grouped Rank Tests with Applications to Screening Experiments," Biometrics 19, 58-84.

14. B.R. Bhat, (1964), "Bayes Solution of Sequential Decision Problem for Markov Dependent Observations," Ann. Math. Statist. 35, 1656-1662.

15. R.A. Bradley, D.C. Martin and F. Wilcoxon (1956) "Sequential Rank Tests: I. Monte Carlo Studies of the Two-Sample Procedure," Technometrics 8, 615-623.

16. R.A. Bradley, S.D. Merchant and F. Wilcoxon, (1966) "Sequential Rank Tests: II Modified Two-Sample Procedures," Technometrics 8, 615-623.

17. Y.T. Chien and K.S. Fu (1966), "A Modified Sequential Recognition Machine Using Time-Varying Stopping Boundaries, " IEEE Trans. Inform. Theory 12, No. 2, 206-214.

18. E.A. Patrick and J.C. Hancock, (1966) "Nonsupervised Sequential Classification and Recognition of Patterns," IEEE Trans. Inform. Theory 12, 362-372.

19. I.R. Savage and J. Sethuraman, (1966) "Stopping Time of a Rank-Order Sequential Probability Ratio Test Based on Lehman Alternatives," Ann. Math. Statist. 37, No. 5, 1154-1160.

20. G.B. Wetherill, (1966) "Sequential Methods in Statistics, Methuen, London and Wiley New York.

21. T.S. Ferguson (1967), Mathematical Statistics. A Decision Theoretic Approach, Ch. 7, Academic Press, New York.

22. K.S. Fu and G.P. Cardillo (1967), "An Optimum Finite Sequential Procedure for Feature Selection and Pattern Classification," IEEE Trans. Auto. Control 12, 588-591.

23. K.S. Fu and Y.T. Chien, (1967) "Sequential Recognition Using a Nonparametric Ranking Procedure," IEEE Trans. Inform. Theory 13, 484-492.

24. R.E. Bechhofer and M. Sobel (1968), Sequential Identification and Ranking Procedures, University of Chicago Press, Chicago and London.

25. R.H. Berie and I.R. Savage (1969) "The Information in a Rank-Order and the Stopping Time of Some Associated SPRT's," Ann. Math. Statist. 39, 1661-1674.

26. K.S. Fu (1970), Sequential Methods in Pattern Recognition and Machine Learning, Academic Press, New York and London.

27. R.C. Milton (1970), Rank Order Probabilities, Ch. 3, John Wiley, New York.

28. K.S. Fu (1970), Sequential Methods in Pattern Recognition and Machine Learning, Academic Press, New York and London.

29. R.C. Milton (1970), Rank Order Probabilities, Ch. 3, John Wiley, New York.

30. M. Zielezny and O.J. Dunn, (1975) "Cost Evaluation of a Two-Stage Classification Procedure," Biometrics 81, 37-47.

31. P. Papantoni-Kazakos, (1976) "Observation Delays in Sequential Estimation," IEEE Trans. Inform. Theory,IT-22(1),January 1976.

32. P. Papantoni-Kazakos, "Bayes Estimation with Asymetric Cost Function," IEEE Trans. Inform. Theory, IT-21 (1), January 1975.

BOOLEAN FLOW THEORY AND ITS APPLICATIONS TO CLUSTER ANALYSIS, FUZZY LOGICS AND PARTICLE TRANSMISSION

KENJI ONAGA[*] and WATARU MAYEDA[**]
[*] Institute of Scientific and Industrial Research, Osaka University
Suita-shi, Osaka, JAPAN
[**] Coordinated Science Laboratory, University of Illinois, Urbana
Illinois, U.S.A.

ABSTRACT

A boolean mapping f from the edge set E of a network $N=(G,I)$, $I=[\alpha,\beta]$ into a power set $\wp(S)$ of a certain set S is called a boolean flow if the continuity condition holds at each node. f is said admissible in N if $\alpha(e) \subseteq f(e) \subseteq \beta(e)$, $\forall e \in E$. Algebraic and topological properties of boolean flows, a nacessary and sufficient condition for existence of admissible boolean flows in N, and characterization of maximum boolean flows are presented in this paper. Also proved are: the terminal capacity of two-terminal boolean flows satisfies the boolean triangular inequality, which leads us to a notion of boolean network metric. γ-relativeness of a symmetric boolean mapping m for any threshold $\gamma \subseteq S$ induces a cluster decomposition of the node set V if and only if m is a boolean network metric.

INTRODUCTION

Transmission of information appears in several different modes such as relay of messages in a communication network, concatenation of propositions in logics, and transmission of similarity in pattern recognition.

Transmission of information can be viewed as a flow of particles through a capacitated network. In the communication network one bit of information is regarded as a particle and edge capacity limits number of passing particles. A flow is arithmetrically additive in a sense that if two flows of f_1 and f_2 bits/sec merge, a single flow of f_1+f_2 bits/sec results. In the logical network truth of a logical proposition is regarded as a particle and edge capacity permits or not permit passage of the particle. The flow is set-additive in a sense that if two flows of f_1 and f_2 (measured by presence =1 or absence=0 of the particle) join together, a single flow of $f_1 \vee f_2$ results. If the particle can pass through the network from input to output, then the compound proposition is regarded as true.

In the above we have informally introduced two natures of information transmission, set-additivity and arithmetic-additivity. In this paper we develope theory of boolean flows in the network where reachability of different kind particles is primarily concerned, while quantity of single kind particles is a problem in the ordinary network flow. We describe in details how to apply boolean flow theory on some aspects of fuzzy set theory and cluster analysis.

BOOLEAN FLOWS IN A NETWORK

Let $G=(V,E)$ be a directed graph with a node set V and edge set E, and let S be a set (of particles). A network $N=(G,I)$ is a directed graph G with each edge e of G weighted by an interval $I(e)=[\alpha(e), \beta(e)]$ of S

$$[\alpha(e), \beta(e)] \triangleq \{A : \alpha(e) \subseteq A \subseteq \beta(e) \subseteq S\}. \qquad (1)$$

A function f on E taking as the value a subset of S, i.e.,mapping $f:E \to \wp(S)$ =collection of all the subsets of S, is called a boolean flow (simply flow if no confusion arises) on G if it satisfies Boolean Kirchhoff's current law

$$\bigcup_{e \text{ in } K^+} f(e) = \bigcup_{e \text{ in } K^-} f(e) \qquad (2)$$

for every cutset $K=K^+ \cup K^-$ of G, where K^+(or K^-) is a set of edges in K whose directions agree (or disagree) with the cutset orientation and f(e) is the value of f on edge e. We freely use in the sequel set theoretic operations union $\cup$, intersection $\cap$, inclusion relation $\subseteq$, element of $\in$, and complement $^{-}$. If flow f on G satisfies the edge admissibility constraint

$$f(e) \in I(e), \text{ i.e., } \alpha(e) \subseteq f(e) \subseteq \beta(e), \ \forall e \in E. \tag{3}$$

it is called admissible on the network N. Let $\mathcal{F}_N$ be the collection of all admissible flows on N and define a maximum flow $\hat{f} \in \mathcal{F}_N$ with respect to a specific $\bar{e}=(\bar{z},z)$ by the condition

$$\hat{f}(\bar{e}) \supseteq f(\bar{e}), \quad \forall f \text{ in } \mathcal{F}_N. \tag{4}$$

[Example] Let $S=\{a,b,c,d\}$, directed graph G be given in Fig.1 and $I=[\alpha,\beta]$ be in Table 1. Consider functions f_1, f_2 defined by Table 1.

(1) f_1 is a flow on G. For cutset $K=\{e_2,e_8\}^+ \cup \{e_3,e_4\}^-$, we can verify Eq.(2).

$$\bigcup_{e \text{ in } K^+} f_1(e) = f_1(e_2) \cup f_1(e_8) = \{a,b,c,d\}$$
$$\bigcup_{e \text{ in } K^-} f_1(e) = f_1(e_3) \cup f_1(e_4) = \{a,b,c,d\}.$$

On the other hand f_2 is not a flow, for it does not satisfy Eq.(2) for the same cutset K

$$\bigcup_{e \text{ in } K^+} f_2(e) = f_2(e_2) \cup f_2(e_8) = \{a,c\}$$
$$\bigcup_{e \text{ in } K^-} f_2(e) = F_2(e_3) \cup f_2(e_4) = \{a,b\}.$$

(2) f_1 is admissible on N and is actually maximum with respect to edge $e=(v_2,v_1)$.

Figures and figure captions

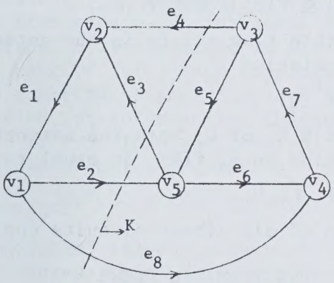


Fig.1 Directed graph G=(V,E) for boolean flow examples

Table 1 Intervals I(e) and functions f_1 and f_2

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8
α	{a}	{a,b}	{a,b,c}	{d}	{c}	{d}	{c,d}	ϕ
β	{a,b,c,d}	{a,b,d}	{a,b,c}	{b,c,d}	{a,b,c}	{c,d}	{a,c,d}	{a,c}
f_1	{a,b,c,d}	{a,b,d}	{a,b,c}	{d}	{c}	{c,d}	{c,d}	{c}
f_2	{a}	{a}	{a}	{b}	{b,c}	{c}	{c}	{c}

Boolean flows on the network N=(G,I) possess fllowing properties whose proof is straight forward, hence omitted (see Onaga [1] for details).

[1] If f_1 and f_2 are flows on G, so are $f_1 \cup f_2$ and $\delta \cap f_1$ $(\delta \subseteq S)$.

[2] On a strongly connected graph any constant function $f=\delta(\delta \subseteq S)$ is a flow but the empty function $f=\phi$ is the only flow possible on an acyclic graph.

[3] For a direct circuit L(edges are all directed in the same orientation), define a function f_L by

$$f_L(e) \triangleq \begin{cases} S & \text{if } e \text{ in } L \\ \phi & \text{if } e \text{ not in } L \end{cases} \qquad (5)$$

Then f_L is a flow on G in the simplest form.

[4] A function f is a flow on G if and only if it has a decomposition of a form

$$f = \bigcup_{L \text{ in } \vec{\mathbb{L}}} [\ \delta_L \cap f_L\ ] \qquad (6)$$

where $\delta_L \subseteq S$, $\vec{\mathbb{L}}$ = collection of all direct circuits of G.

[5] Let f_1 and f_2 be admissible. Then $\delta \cap f_1$, $\delta \supseteq \bigcup_{e \text{ in } E} \alpha(e)$ and $(\delta_1 \cap f_1) \cup (\delta_2 \cap f_2)$, $\delta_1 \cup \delta_2 = S$ are also admissible in N.

ADMISSIBLE FLOWS AND MAXIMUM FLOWS

In this section we derive a necessary and sufficient condition for existence of admissible flows in the netsork N=(G,I). We begin by a simple observation. For the sake of notational simplicity, we use

$$\begin{aligned} &f(A,\cup) = \bigcup_{e \text{ in } A} f(e), \quad f(A,\cap) = \bigcap_{e \text{ in } A} f(e) \\ &f(A,\cup;\mathbb{A},\cap) = \bigcap_{A \text{ in } \mathbb{A}} [\ \bigcup_{e \text{ in } A} f(e)] \end{aligned} \qquad (7)$$

[6] Let f be an admissible flow on N. Then $f(\bar{e})$ satisfies

$$f(\bar{e}) = f(L,\cap;\ \vec{\mathbb{L}}_0,\cup) \qquad (8)$$

$$\alpha(L,\cap;\vec{\mathbb{L}}_0,\ \cup) \subseteq f(\bar{e}) \subseteq \beta(L,\cap;\vec{\mathbb{L}}_0,\ \cup) \qquad (9)$$

[Theorem 1] An admissible flow exists in the network $N=(G,I=[\alpha,\beta])$ if and only if an inclusion relation

$$\alpha(K^-,\cup) \subseteq \beta(K^+,\cup) \qquad (10)$$

holds for every cutset $K=K^+ \cup K^-$ of G. When the network N permits admissible flows, the maximum flow value on $\bar{e}$, $\hat{f}(\bar{e})$, is equal to

$$\hat{f}(\bar{e}) = \beta(L,\cap;\ \vec{\mathbb{L}}_0,\cup) \qquad (11)$$

where $\vec{\mathbb{L}}_0$ is the collection of all direct circuits containing $\bar{e}$.

TWO-TERMINAL FLOWS

Let node z and $\bar{z}$ be specified as the input and output of the network N, respectively. Ajoin the terminal edge $\bar{e}=(\bar{z},z)$, $\alpha(\bar{e})=\phi$, $\beta(\bar{e})=S$ to form a new network N* and let f* be an admissible flow in N*. Then the restriction f of f* to the original network N is called an admissible two-terminal flow in N and $f^*(\bar{e})$ is called the terminal flow of f and denoted by F. The maximum of F is called the terminal capacity of N

$$\hat{F} = \mathrm{Max}_{f^* \text{ in } \mathcal{F}_{N^*}} f^*(\bar{e}) \qquad (12)$$

The terminal capacity expresses reachability of particles in the network, that is, the collection of particles that can be transported through the network from node z to $\bar{z}$. The set of various terminal capacities between different input-output pairs is a good measure of communication capability of the network. To distinguish the input-output pair (x,y), we use notations of x→y flow and x→y capacity $\hat{F}_{xy}$.

[Theorem 2] (a) An admissible x→y flow exists in N if and only if

$$\alpha(K^-,\cup) \subseteq \beta(K^+,\cup)$$

holds for every cutset $K=K^+ \cup K^-$ of G except y→x cutsets.

(b) x→y capacity $\hat{F}_{xy}$ is given by

$$\hat{F}_{xy} = \begin{cases} \beta(Q, \cap;\ \vec{Q}_{xy}, \cup) \\ \beta(J, \cap;\ \vec{\mathbb{J}}_{xy}, \cup) \end{cases} \tag{13}$$

where $\vec{Q}_{xy}$ is the collection of all x→y direct paths of G and $\vec{\mathbb{J}}_{xy}$ the collection of all x→y direct edge sequences of G.

(c) The terminal capacity satisfies the following recursive equations.

$$\hat{F}_{xy} = \begin{cases} \bigcup_{w \in V} [\hat{F}_{xw} \cap \beta(e) : e=(w,y) \in E] & (14) \\ \bigcup_{w \in V} [\hat{F}_{xw} \cap \hat{F}_{wy}] & (15) \end{cases}$$

(d) The terminal capacity satisfies the boolean triangular inequality

$$\hat{F}_{xz} \supseteq \hat{F}_{xy} \cap \hat{F}_{yz} \quad \text{for all } x,y,z \text{ in } V$$

where $\hat{F}_{xx} \triangleq S$ for every x in V.

Theorem 2 provides a justification of the following node labelling method for the terminal capacity calculation.

≪Calculation of the terminal capacity $\hat{F}_{zx}$, $\forall x \in V$≫

1° Set $\sigma(z)=S$, $\sigma(x)=\phi$, $\forall x \neq z$.

2° Modify $\sigma(x)$ by

$$\sigma(x) \leftarrow \sigma(x) \cup \{\sigma(y) \cap \beta(e) : e=(y,x) \in E\} \tag{16}$$

3° Stop if no change occurs. Then $\sigma(x)= \hat{F}_{zx}$.

Modification 2° should be, in general, carried out in the order of increasing edge-distance from node z. In this case $\sigma(x)$ can be proved to converge at most $|V|-1$ sweeps. $\sigma(x)$ converges in one sweep if G is acyclic and if the modification is carried out in the topological order of the nodes, in which the nodes are numbered so that $(i,j) \in E$ implies $i<j$.

[Theorem 3]* (Max-flow cut-intersection Theorem) The value of a cutset $K=K^+ \cup K^-$ is defined by $(K^+,\cup)$. Then x→y terminal capacity $\hat{F}_{xy}$ is equal to the intersection of values of all x→y cutsets

$$\hat{F}_{xy} = \bigcap_{K \in \mathbb{K}_{xy}} \beta(K^+,\cup) = \beta(K^+, \cup;\ \mathbb{K}_{xy}, \cap) \tag{17}$$

where $\mathbb{K}_{xy}$ is the collection of all x→y cutsets of G.

Eq. 17 is the dual expression of Eq.13 and can be viewed as a generalization of the max-flow min-cut theorem of Ford-Fulkerson [3], but we should bear in mind that in the boolean case there is no min-cut whose value is equal to the max-flow.

TRANSMISSION OF PATTERN FEATURES IN SIMILARITY EXPERIMENT

Consider a similarity experiment on n patterns, in which a pattern x is characterized by p(x), a subset of the pattern feature space P. Similarity S_{xy} between x and y is defined to be a set of features common to both, $S_{xy} \triangleq p(x) \cap p(y)$ and dissimilarity D_{xy} is equal to the complement of S_{xy} with respect to P, $D_{xy} \triangleq \bar{S}_{xy}$. For any three patterns x,y,z, a boolean triangular inequalities of two kinds

$$S_{xz} \supseteq S_{xy} \cap S_{yz} \tag{18}$$

$$D_{xz} \subseteq D_{xy} \cup D_{yz} \tag{19}$$

are readily seen to hold.

---

* This is a specialization of the well-known distributive law for distributive lattices and a generalization of Edmonds-Fulkerson [2] for real numbers.

In practical situations, direct measurement of pattern feature p(x) is not possible but similarity between two patterns S_{xy} may be feasible. several tests should be employed, for one test is capable of measuring only a subset of true S_{xy}. An other important factor to consider is accuracy and asymmetry of measurement. It is rather safe to assume that tests only reveal an optimistic estimate β_{xy} and a pessimistic estimate α_{xy} of the true similarity

$$\alpha_{xy} \subseteq S_{xy} \subseteq \beta_{xy} \tag{20}$$

and may be asymmetric if it favors one pattern to the other. We construct a boolean network $N=(G,I=[\alpha,\beta])$: Represent pattern x by node x and draw a direct edge $e=(x,y)$ if an asymmetric test e of pattern x with respect to pattern y is conducted. Edge $e=(x,y)$ is weighted by an interval $I(e)=[\alpha(e),\beta(e)]$ of the space P.

With those data in hand, what can we say about the similarity of pattern x with respect to pattern y ? In order for pattern x and y to share a feature s, a point of P, in common, feature s must be transported through the network from node x to y, in order words, there must exist x→y directed path Q such that $s \in \beta(Q,\cap)$. This suggests the similarity S_{xy} is a collection of pattern features transportable through the network from x to y. This is no other than x→y flow in N. This x→y flow must be admissible in N in order to be usefull and consistent with the framework of the experiment.

$$\alpha(e) \subseteq f(e) \subseteq \beta(e), \forall e \in E \tag{21}$$

The value f(e) represents a probable outcome of test e and the terminal flow F_{xy} is regarded as an estimate of S_{xy} accounting all combination of test sequences. The terminal capacity $\hat{F}_{xy}$ is now interpreted as an optimistic estimate of S_{xy} deducible from data accumulated.

[Theorem 4] Let the network be symmetric and let S_{xy} be the similarity between x and y derived from the network. Pattern x and y are called γ-relative ($\gamma \subseteq P$) if $S_{xy} \supseteq \gamma$. Then γ-relativeness is an equivalence relation and dissimilarity numerical measure $d_{xy} \triangleq \mu(\bar{S}_{xy})=1-\mu(S_{xy})$ is an Euclidean distance.

It is well known in algebra that an equivalence relation induces a partition of the node set V, $V=\bigcup_j V_j$, $V_i \cap V_j=\phi$, $i \neq j$, such that every node pair x and y in V_i is γ-relative. For example consider γ-relativeness of $V=\{x_1,x_2,\cdots,x_{10}\}$ for a decreasing threshold sequence $P \supset \gamma_1 \supset \gamma_2 \cdots \supset \gamma_6$

P-relative: $\{x_1\}\ \{x_2\} \cdots\cdots \{x_{10}\}$
γ_1-relative: $\{x_1,x_2,x_3\}\ \{x_4\}\ \{x_5\} \cdots\cdots \{x_{10}\}$
γ_2-relative: $\{x_1,x_2,x_3\}\ \{x_4\}\ \{x_5\} \cdots\cdots \{x_8\}\ \{x_9,x_{10}\}$
γ_3-relative: $\{x_1,x_2,x_3\}\ \{x_4,x_5\}\ \{x_6\}\ \{x_7,x_8\}\ \{x_9,x_{10}\}$
γ_4-relative: $\{x_1,x_2,\cdots,x_6\}\ \{x_7,x_8\}\ \{x_9,x_{10}\}$
γ_5-relative: $\{x_1,x_2,\cdots,x_6\}\ \{x_7,x_8,\cdots,x_{10}\}$
γ_6-relative: $\{x_1,x_2,\cdots,x_{10}\}$

The hierarchy of γ-relativeness is best visualized by the hierarchy tree of Fig.2. A cluster C of V is usually defined as a set of nodes such that the intra-set similarity (or distance) is larger (or smaller) than the interset one,

$$x,y \in C, \quad z \notin C \quad \text{implies} \quad S_{xy} \supsetneqq S_{xz}\ ,\ S_{yz} \tag{22}$$

$$\text{or} \quad d_{xy} < d_{xz}\ ,\ d_{yz} \tag{23}$$

Each component of the partition of V due to the γ-relativeness clearly satisfies Eqs.22 and 23 and hence forms a cluster. We will prove later the converse of the above that if a system of clustering of V forms a hierarch tree, the similarity S_{xy} used in Eq.22 must satisfy the boolean triangular inequality of Eq.18.

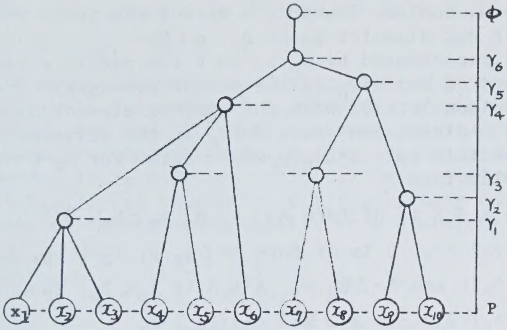


Fig.2 Hierarchy tree induced by γ-relativeness for an increasing sequence of thresholds γ

TRANSMISSION OF FUZZY LOGIC PROPOSITIONS

Person x and y are called 0-friend if they are friend, and 1-friend if they are friends or have at least one common friend. Consider a problem of investigating social communication of n persons $V=\{x_1,x_2,\cdots,x_n\}$ randomly sampled from a certain society. 1-friendship rather than 0-friendship should be investigated because it is likely few of these persons are 0-friend each other, and social communication via two or more intermediate friends is weak and unreliable. F_x denoting the set of friends of x, $F_{xy}= F_x \cap F_y$ expresses common friends of x and y. x and y are 1-friend if F_{xy} is nonempty. In practical situations it is quite reasonable to assume a pessimistic estimate $\alpha(x,y)$ and optimistic estimate $\beta(x,y)$ of F_{xy}, $\alpha(x,y) \subsetneq F_{xy} \subseteq \beta(x,y)$ are more readily measurable. Different methods of measurement are employed and a symmetric boolean network N=(G,I) is constructed as described in Section 5.

A x-y two-terminal boolean flow admissible in N is a realization of 1-friendship among those persons with regard to x and y, and x-y terminal flow F_{xy} a set of common friends of x and y. It is now clear that transmission of logical proposition "- and - are 1-friend" can be faithfully modeled by reachability of different kind particles through a boolean network.

Suppose person x and y are asked if person z is a friend of his and that their answers are, "I(x) know him. We once in a while exchange hello on the street." and "We(y and z) are old friend from high school days." Shall we decide x and y are 1-friend ? This question raises difficulty inherent in any set-theoretic approach to problems involving human subjectivity or ambiguity. Fuzzy set theory was proposed in 1965 by Zadeh [9] to surmount this difficulty by introducing "degree of friendliness" : in the above we should rather have asked "What is your degree of friendliness with person z ?" Suppose $f^{(0)}_{xz}=0.6$, $f^{(0)}_{yz}=0.9$ are their answers. The degree of 1-friendliness $f^{(1)}_{xy}$ between x and y is naturally set to

$$f^{(1)}_{xy} = \mathrm{Min}\ [\ f^{(0)}_{xz}\ ,\ f^{(0)}_{yz}\ ] = \mathrm{Min}\ [0.6,\ 0.9] = 0.6$$

if we regard "friendship with z" as the particle and "degree of friendliness" as its diameter. Fix z and measure pessimistic and optimistic estimates between pairs (x,y). We construct a network N=(G,I) as usual. 1-friendship between x and y with z as a reference is faithfully modeled by reachability of single kind particles with various diameters. The particle is a single kind because z is a reference and the diameter corresponds to "degree of 1-friendliness". For any person in the society can be a reference, we are led to consider reachability of different kind particles with differing diameters, boolean flows in a direct sum space $S=\Sigma_{p\in P} D_p$ of particle-kind space P and diameter space D_p, $p\in P$.

A particle is represented by $d_p.p$, $p\in P$ the particle kind and d_p its diameter. The diameter space D_p is the set of nonnegative real numbers or in more general a boolean lattice with the minimum element 0, maximum element M and ordering $\leqslant$. The direct sum space $S=\Sigma D_p$ is the universal set of our concern: space of particle sets $s=\Sigma d_p.p$, where terms for $d_p=0$ are omitted. We observe the followings.

(a) A subset $A\subseteq S$ is of form $A=\Sigma A_p.p$, $A_p\subseteq D_p$.

(b) An interval I of S is of form $I=\Sigma I_p.p$, $I_p=[\alpha_p, \beta_p]$ interval of D_p

(c) Let $A=\Sigma A_p.p$ and $B=\Sigma B_p.p$. $A\subseteq B$ if $A_p\subseteq B_p$, $\forall p\in P$. $\bar{A}=\Sigma \bar{A}_p.p$. $A\cap B=\Sigma (A_p\cap B_p)\cdot p$. $A\cup B=\Sigma (A_p\cup B_p)\cdot P$.

(d) A boolean flow f is of form $f=\Sigma f_p.p$, $f_p\subseteq D_p$ and the boolean Kirchhoff's current law on f implies that on f_p and conversely.

It is now clear that an admissible boolean flow in N=(G,I) is a direct sum of admissible boolean flows f_p in $N_p=(G,I_p)$. A comment is in order at this point. The diameter space D_p is allowed to be a boolean lattice while we have been dealing only with set-lattices. This raises no problem because there is a theorem in algebra that any boolean lattice is isomorphically represented by a set-lattice, collection of subsets of a some set closed under $\cap, \cup, {}^{-}$,[14].

NETWORK METRICS

There is insatiable desire in our society for ranking, ordering and clustering objects such as seen in nomination of a person for a prize, decision of an economic policy among alternatives and classification of animals or books into categories. For this purpose notion of metrics is essential.

Define a boolean metric m on $N=(G,I=[\alpha,\beta])$ as a mapping from VxV to $\mathcal{P}(S)$ satisfying the boolean triangular inequality

$$\begin{cases} m(x,x)=S, \ \forall x\in V \\ m(x,z)\supseteq m(x,y)\cap m(y,z), \ \forall x,y,z\in V \end{cases} \tag{24}$$

(The dual form $m(x,z)\subseteq m(x,y)\cup m(y,z)$ can alternatively be used.)

A boolean metric m is said admissible in N if

$$m(x,y)\in I(e) \text{ or } \alpha(e)\subseteq m(x,y)\subseteq\beta(e), \ \forall e=(x,y)\in E \tag{25}$$

We often write m(e) for m(x,y) if (x,y) is edge e. Metric m is called maximum or minimum in N if $m\supseteq m'$ or $m'\subseteq m$ holds for any admissible m' in N.

[Theorem 5] (a) $\delta\cup m$, $m_1\cap m_2$ are metrics if $\delta\subseteq S$ and m, m_1, m_2 are metrics.

(b) $(\delta_1\cup m_1)\cap(\delta_2\cup m_2)$ is admissible in N if $\delta_1\cup\delta_2=S$, $\delta_1\cap\delta_2=\phi$ and m_1, m_2 are admissible metrics in N.

(c) Define h_X for a subset X of V by

$$h_X(x,y)\triangleq\begin{cases}\phi & \text{if } x\in X,\ y\in\bar{X} \\ S & \text{otherwise}\end{cases} \tag{26}$$

A mapping $m: V\times V\rightarrow\mathcal{P}(S)$ is a metric if and only if it has a decomposition of a form

$$m=\bigcap_{i=1}^{k}[\delta_i\cup h_{X_i}] \tag{27}$$

(d) Define m_α by

$$m_\alpha(x,y)=\alpha(Q,\cap;\vec{Q}_{xy},\cup) \tag{28}$$

Then an admissible metric exists in the network N=(G,I) if and only if

$$m_\alpha(e) \subseteq \beta(e), \quad \forall e \in E \tag{29}$$

In this case m_α is actually a minimum metric admissible in N.

[Theorem 6] The γ-relativeness of a symmetric metric m induces a cluster decomposition Π of V. For any increasing sequence $\phi = \gamma_1 \subset \gamma_2 \subset \cdots \subset \gamma_k = S$, the corresponding cluster decomposition $\pi_1, \pi_2, \cdots, \pi_k$ constitute a hierarchy tree, in other words, if γ_i-relative set $C(\pi_i)$ is partitioned by γ_{i+1}-relativeness into $C_1(\pi_{i+1}), C_2(\pi_{i+1}), \cdots, C_p(\pi_{i+1})$, then form a subtree of Fig.3. Continuation of this process results in a hierarchy tree.

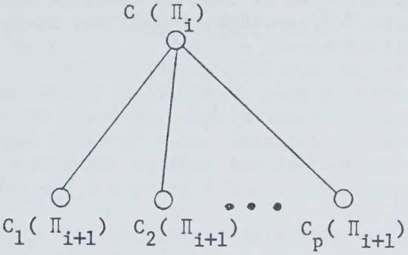


Fig.3 Construction of hierarchy tree from a cluster decomposition

CONCLUSIONS

Upon a basic premise that the carrier of information is a particle and information is either set-additive or scalar-additive, we have developed a theory of boolean flows in a network N in hope of providing a unified theory of transmission of information such as messages, similarities, fuzzy logics, and clustering. The terminal capacity is shown to be a good measure for expressing particle transmission capability of the network. Moreover the terminal capacity is shown to satisfy the boolean triangular inequality and leads us to the notion of boolean metrics. The symmetric metric plays an essential role for clustering objects: γ-relativeness for a threshold $\gamma \subseteq S$ is proved to induce a cluster decomposition of V. For any increasing sequence of thresholds, a hierarchical tree can be constructed to show the structure of the objects. Conversely any boolean symmetric mapping inducing a cluster decomposition of V must be a metric.

It seems to the authors that we have succeeded in showing notion of boolean flows plays a unifying role in describing, understanding, and analyzing transmission of various kind information through a series of detailed examples and usefull theorems. Moreover we have proposed the boolean metric as the axiom of information transmission and showed its utility in cluster analysis.

In this paper we have defined the dissimilarity by $D_{xy} = \bar{P}_x \cup \bar{P}_y$, but there is a more plausible alternative definition, Hamming dissimilarity $D_{xy} = P_x \oplus P_y$, $\oplus$ = symmetric difference or ring sum. This comment forces us to distinguish between boolean algebra flows and boolean ring flows. Is it possible to construct a more general theory of boolean flows, boolean tensions and boolean potentials in the network ? A preliminary version is seen in Onaga [1,6].

This work was supported in part by the Grant in Aid for Scientific Research of the MInistry of Education, Science and Culture of Japan under Grant : Cooperative Research (A) 135017 (1976).

REFERENCES

[1] K.Onaga,"Properties of boolean flows in networks," Trans.IECE, Japan, Vol.59-A, No.2, pp.92-99, 1976.

[2] J.Edmonds and D.R.Fulkerson, "Bottleneck extrema," Journal of Combinatorial Theory, Vol.8, pp.299-306, 1970.

[3] L.R.Ford,Jr. and D.R.Fulkerson, Flows in networks, Princeton University Press, Princeton, N.J., 1962.

[4] L.A.Zadeh, "Fuzzy sets," Infor.Contro., Vol.8, pp.338-353, 1965.

[5] S.Iyanaga and K.Kodaira, Introduction to Modern Mathematics, Iwanami-shoten Tokyo, 1967.

[6] K.Onaga, "Basic properties of boolean tensions and boolean potentials in boolean networks," Trans.IECE, Japan, Vol.59-A, No.2, pp.100-107, 1976.

A NUMERICAL METHOD FOR THE COMPUTATION OF PROBABILITY OF ERROR IN A STRUCTURED DETECTOR OF AN OPTICAL COMMUNICATION SYSTEM

P. R. KUMAR
Dept. of Systems Science and Mathematics
Washington University
St. Louis, Missouri 63130

LESTER F. EASTWOOD, JR., Ph.D.
Center for Development Technology and
Dept. of Electrical Engineering
Washington University
St. Louis, Missouri 63130

ABSTRACT

The increasing use in optical communication of avalanche diodes for detection has made it necessary to compute the probability of error in such structured detectors. Previous efforts have been directed either towards finding upper bounds or approximating such probabilities. Here we perform a numerical evaluation of the exact expression for the error rate and compare the results with those of other techniques. The results of this paper yield a correct value for optimum system gain and are therefore of significant interest in the design of fiber optics receivers.

I. THE COMMUNICATION SYSTEM

The structure of a typical binary optical communication system is shown in Figure 1. The transmitter is controlled by a data source emitting a sequence of "ones" and "zeroes." In each of the two states "zero" and "one", the transmitter emits a pulse of optical energy of fixed duration, say T seconds, but of one of two possible intensity levels.

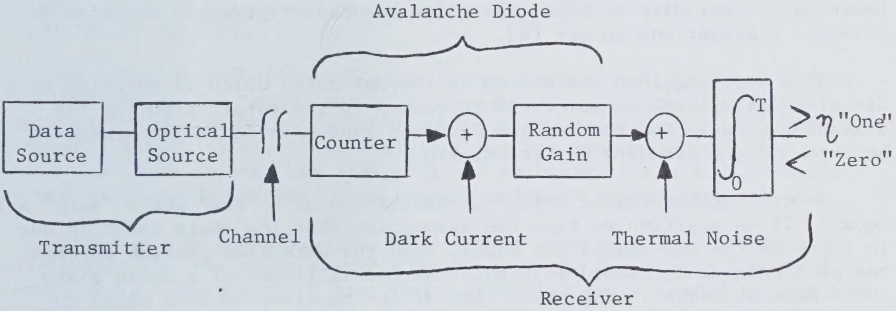


Figure 1. Optical Communication System

At the receiver the optical energy passes through an optical fiber and impinges on an avalanche photodiode. The optical pulse of T seconds duration causes the emission of photo-electrons which constitute the "primary" current. Each of these primary photo-electrons moves through the diode and is multiplied by the avalanche process occurring in the diode. The output current is integrated for the time period T seconds, and based on the output of the integrator at the end of this time period (or equivalently on the total number of output counts) a decision of either a "one" or a "zero" is to be made.

II. THE MODEL

The number of primary photoelectrons emitted in the T-second time period is modeled as a Poisson-distributed random variable with parameter dependent on the intensities of the incoming optical energy and the background energy. When the signal is a "one," the parameter of the Poisson-distributed random variable is [LAMD + LAMSIG] where LAMSIG is the contribution of the signal and LAMD that of the background (or "dark current"). In the "zero" state the parameter is [LAMD + EXT•LAMSIG], where EXT denotes the extinction ratio, ideally zero, but nonzero in practical systems.

Statistical modeling of the random gain mechanism of the avalanche diode is more complex. Using physical arguments, Personick [1] obtained differential equations for the moment generating functions involved. By utilizing this technique, he was able to calculate the Chernoff bound for the error rate of the receiver.

An alternative approach has been suggested by McIntyre [2]. By considering various special cases of the problem he conjectured an expression for the probability of obtaining (N+R) output electrons with N primary electrons as input:

$$P_{N,N+R} = \frac{N(1-K)^{R}}{(N+KR)\ R!}\ \frac{\Gamma\left(\frac{N+R}{1-K}\right)}{\Gamma\left(\frac{N+KR}{1-K}\right)}\ \left[\frac{1+K(M-1)}{M}\right]^{\left(\frac{N+KR}{1-K}\right)} \left[\frac{M-1}{M}\right]^{R} \qquad (1)$$

There was reason to believe the validity of this expression since it could be shown to be true for the case of K (a parameter of the diode) equal to 0·5. Moreover it has been experimentally verified by Conradi [3]. Recently the validity of this expression has been rigorously verified by Balaban, Fleisher and Zucker [4].

Also entering into the system is thermal noise which is modelled as a Gaussian distributed random variable adding to the total output of the avalanche diode. The mean of the Gaussian random variable is assumed to be zero and its standard deviation, SIG.

We will assume that a priori probabilities of a "one" and a "zero" are equal. If in addition we make the assumption that the costs incurred due to false alarms and misses are equal, then the detection problem becomes one of finding a threshold such that the probabilities of a false alarm and a miss are equal. The error rate of the receiver is then equal to either of these quantities.

III. BACKGROUND

The above problem has arisen from the need to analyze the performance of optical communication circuitry in practical situations. The avalanche diodes, which constitute an integral part of the optical communication systems used for detection, lie at the heart of the problem's mathematical intractability. The expression defining the random gain of the diode is very complex and so consequently is the computation of the error rate.

Personick [1] was the first to consider the above problem. By analyzing the system in the frequency domain he was able to obtain the Chernoff bound for the error rate of the receiver. However the relative tightness of the Chernoff bound cannot be determined, and this constitutes the major disadvantage of that method.

Concurrently with the work described in this paper, other efforts have been made to analyze receiver performance. However, these techniques have not utilized the exact expression for the random gain of the avalanche diode and only obtain approximations for the probability of error. One such effort by Balaban makes use of an expression suggested by McIntyre [2] to approximate the random gain of the avalanche diode. For a detailed comparison of the technique described in this paper and other techniques, see [5].

In this paper we show that calculations utilizing this exact expression are indeed feasible and outline a numerical method for this purpose. We also present the results of this technique and compare the data obtained with the results of the Chernoff bounding technique of Personick [1] and the approximation technique of Balaban [5]. The results can similarly be used to judge the performance of other techniques to approximate the error rate of the optical communication system described above.

IV. PROBABILITY OF ERROR COMPUTATION

We have devised a straightforward numerical procedure to calculate the system probability of error. The probability distributions of the input counts to the avalanche diode in the two information states are evaluated and the probability distributions of the outputs computed. The output distribution is numerically convolved with a Gaussian distribution to obtain the probability distributions in the two states. A search technique is used to choose a threshold which equates the false alarm and miss probabilities and the total probability of error is calculated. The computational scheme is efficient at those values of K for which $\frac{1}{1-K}$ is an integer. These K values are useful [2]. The error computation can be made as precise as desired.

The input distributions in each of the states "one" and "zero," (the distributions of the random variable representing the number of input counts to the avalanche multiplication process) are evaluated at a number of equally spaced points. The distributions are evaluated only at a finite number of points, the end-points of which are the values at which the probabilities associated with random variables become negligible.

Utilizing now the avalanche diode random gain equation (1) and the distributions of the input random variables in the two states "one" and "zero," the distributions of the random variables representing the output of avalanche diode in the two states are computed. As before, the distributions are only evaluated at points where the associated probabilities are not too small to be neglected. It is essential to this computation to evaluate the random gain expression (1) for a variety of values of N and N+R, and this can be efficiently done only for those values of the parameter K of the diode at which the quantity $\frac{1}{1-K}$ is an integer. For these values of K the Gamma - functions involved become factorials and the expression is efficiently computed by the choice of an appropriate recursive scheme. The recursive scheme develops from the fact that we can write

$$\log P_{N,N+R} = \log\left[\frac{N}{N+KR}\right] + \sum_{I=1}^{R} \log\left[\frac{\left(\frac{N+R}{1-K}\right)-I}{I} \; \frac{(1-K)(M-1)}{M} \left(\frac{1+KM-K}{M}\right)^{\left(\frac{N+KR}{R-KR}\right)}\right] \tag{2}$$

An initial choice of N and R is made and the above expression is evaluated by a DO loop with value of the index ranging from 1 to R. It becomes computationally infeasible to re-evaluate this expression for other values of input and output counts since R is typically of the order of 10^5. However the random gain expression $P_{N+J,N+R}$ for a perturbation J in the input can be written

$$\frac{P_{N+J,N+R}}{P_{N,N+R}} = \frac{(N+J)(N+KR)}{N(N+J+KR-KJ)} \, \frac{R!}{(R-J)!} \, \frac{\Gamma\left[\frac{N+KR}{1-K}\right]}{\Gamma\left[\frac{N+KR}{1-K}+J\right]} \left[\frac{1+K(M-1)}{(1-K)(M-1)}\right]^{J} \tag{3}$$

Hence

$$\log P_{N+J,N+R} = \log P_{N,N+R} + \log \frac{(N+J)(N+KR)}{N(N+J+KR-KJ)}$$
$$+ \sum_{I=1}^{J} \log\left[\frac{(R-J+I)}{\left(\frac{N+KR}{1-K}\right) - 1+I} \left[\frac{1+K(M-1)}{(1-K)(M-1)}\right]\right] \tag{4}$$

This recursive relationship enables efficient computation of $P_{N+J,N+R}$ for various values of J and so

$$\Pr\,[\text{Avalanche Diode Output} = N+R]$$

$$= \sum_{J=-N}^{\infty} \Pr\,[\text{Input Poisson Random Variable} = N+J]\; P_{N+J,N+R} \tag{5}$$

can be evaluated.

Similarly Pr [Avalanche Diode Output = N+R+L] can also be computed for various values of L. However the quantity $P_{N,N+R+L}$ necessary for this computation need not be calculated afresh but can be obtained from the already computed value $P_{N,N+R}$ by the relationship

$$\log P_{N,N+R+L} = \log P_{N,N+R} + \log\left(\frac{N+KR}{N+KR+KL}\right)$$
$$+ \sum_{J=1}^{L} \log\left[\frac{\left[\frac{N+R}{1-K}+(I-1)\,\frac{1}{1-K}\right]\cdots\left[\frac{N+R}{1-K}+(I-1)\,\frac{1}{1-K}+\frac{1}{1-K}-1\right]}{\left[\frac{N+KR}{1-K}+(I-1)\left(\frac{K}{1-K}\right)\right]\cdots\left[\frac{N+KR}{1-K}+(I-1)\left(\frac{K}{1-K}\right)+\left(\frac{K}{1-K}\right)-1\right]}\right.$$
$$\left.\cdot\left[\frac{(1-K)(M-1)}{M}\cdot\left(\frac{1+K(M-1)}{M}\right)^{\frac{K}{1-K}}\right]\cdot\frac{1}{(R+I)}\right] \tag{6}$$

In this way the probability distribution of the random variable representing the output of the avalanche diode can be computed. However this scheme requires that the quantity $\frac{K}{1-K}$ be an integer, and this constitutes a drawback of this scheme.

To compute the distributions of the total output of the detector it remains to convolve the distributions of the output of the avalanche diode in the two possible states "one" and "zero" with the Gaussian random variable representing the thermal noise in the system. This convolution is performed numerically.

Using a search technique, a threshold level is then located at which the probabilities associated with the "tails" of the distributions in the two states "one" and "zero" are equal. The error rate of the receiver is then equal to either of these quantities.

The above computation can be made as precise as required by increasing the number of interpolation points at which the distributions are computed. However, there is also a corresponding rapid increase in the computational time involved with every increase in the number of interpolation points.

V. RESULTS

The above computation was performed for a given choice of the operating parameters of the system. The error rate of the system was constrained to be equal to 10^{-9} and the signal strengths necessary to achieve this probability of error were computed for various values of the mean gain of the avalanche diode M. The results are compared to those obtained by Personick [1] by constraining the Chernoff bound of the error rate to 10^{-9}, and to the results obtained by Balaban. Results of the Chernoff bounding technique are currently not available for high values of mean gain. As discussed earlier, without any loss of generality, the a priori probabilities of a "one" and a "zero" are assumed to be equal and so are the costs incurred due to false alarms and misses. The results are shown in Table 1.

Table 1: Comparison of Results of Alternative Calculation Methods

System Parameter Values

LAMD = 5
EXT = 0.01
K = 0.5
SIG = 6000
Probability of Error = 10^{-9}

	RESULTS OF THIS METHOD		RESULTS OF CHERNOFF BOUNDING TECHNIQUE OF PERSONICK		RESULTS OF APPROXIMATION TECHNIQUE OF BALABAN	
Mean Gain M	Signal Strength (LAMSIG)	Threshold	Signal Strength (LAMSIG)	Threshold	Signal Strength (LAMSIG)	Threshold
35	2619	38140	2867	42370	2619	38510
40	2433	39620	2772	47120	2450	40300
45	2358	43370			2360	43610
50	2352	48830			2363	49080
60	2452	63490			2453	63100

These results show that the required signal strength for a specified error rate is not very sensitive to the gain of the diode in the neighborhood of the optimal gain. Since the optimal gain is thus obtained as the minimum of a shallow "trough", an accurate calculation of the optimal gain requires an accurate knowledge of the signal strengths required for various values of the gain, which is not obtained by the use of the Chernoff bounding technique. Consequently the Chernoff bounding technique (using values of the Chernoff bound for lower values of mean gain than shown here) yields an optimal gain between 35 and 40 for an error rate of 10^{-9}. The results of this paper, however indicate an optimal gain in the neighborhood of 50, and this fact is of great importance in the design of fiber optics receivers.

VI. SUMMARY

A numerical procedure has been outlined which utilizes the exact expression defining the random gain of the avalanche diode in computing the error rate of the associated communication system. Results obtained by such a computational scheme compare favorably with both earlier and concurrent results. Also an accurate value of the optimal gain is obtained, which is very useful in the design of fibre-optics receiver systems.

VII. ACKNOWLEDGEMENTS

The authors wish to express their gratitude to D. L. Snyder of Washington University and S. D. Personick of Bell Laboratories for many valuable discussions held with them. This research was supported in part by the National Science Foundation under Grant Nos. ENG74-0800 and ENG75-09755.

VIII. REFERENCES

1. S. D. Personick, "Statistics of a General Class of Avalanche Detectors with Applications to Optical Communication," Bell System Technical Journal, Vol. 50, pp. 3075-3095, December 1971.

2. R. J. McIntyre, "The Distribution of Gains in Uniformly Multiplying Avalanche Photodiodes: Theory," IEEE Transactions on Electron Devices, Vol. ED-19, pp. 703-712, June 1972.

3. S. Conradi, "The Distribution of Gains in Uniformly Multiplying Avalanche Photodiodes: Experimental," IEEE Transactions on Electron Devices, Vol. ED-19, pp. 713-718, June 1972.

4. P. Balaban, P. E. Fleisher and H. Zucker, "The Probability Distribution of Gains in Avalanche Photodiodes," to be published.

5. S. D. Personick, P. Balaban, J. H. Bobsin and P. R. Kumar, "A Detailed Comparison of Four Approaches to the Calculation of Error Rates of Optical Fiber System Receivers," to be published in IEEE Transactions on Communication.

OPTIMAL CHOICE OF THE KERNEL FUNCTION FOR KERNEL TYPE DENSITY ESTIMATORS

DIMITRI KAZAKOS
Electrical Engineering Department
State University of New York at Buffalo, Amherst, New York

ABSTRACT

Using an upper bound to the pointwise mean square error performance of the kernel density estimate, we obtain the optimal kernel function. Its expression is:

$Q(y)=(1+a^{-1})\ (2T)^{-1}\ [1-T^{-a}|y|^{a}]$, for $|y|\leq T$, and $a=2-p^{-1}$, $p\geq 1$.

INTRODUCTION

Several probability density function estimators have been proposed as an alternative to the histogram, in order to achieve improved properties. The basic drawback of the histogram is its lack of smoothness. Rosenblatt [1] and Parzen [2] employed the idea of data smoothing in the density estimates they proposed.

A kernel density estimate $\hat{f}_n(x)$ based on a sequence $x_1, \ldots, x_n$ of statistically independent and identically distributed random variables is defined as

$$\hat{f}_n(x) = \int_{-\infty}^{+\infty} K_n(x,y)\, dF_n(y) = n^{-1}\sum_{i=1}^{n} K_n(x,x_i) \tag{1}$$

where $f_n(y)$ is the empirical distribution function based on n observations and has the staircase form. Therefore, $K_n(x,y)$ has a smoothing effect. We will consider kernels K_n of the type

$$K_n(x,y) = h_n^{-1}\, K\, (h_n^{-1}(x-y)) \tag{2}$$

Parzen [2] showed that if K (y) is nonnegative, even around zero and

$$\sup_{-\infty<y<\infty} K(y) < \infty\ ,\ \int_{-\infty}^{+\infty} K(y)\, dy < \infty\ ,\ \lim_{|x|\to\infty} |x|\, K(x)=0 \tag{3}$$

then the asymptotic variance of $\hat{f}_n(x)$ is

$$(n\, h_n)^{-1}\, f(x) \int_{-\infty}^{+\infty} K^2(y)\, dy \tag{4}$$

at the continuity points of the true density function f(x).

The problem of optimizing K(y) was first studied by Epanechnikov [3]. He restricted the class of kernel functions by imposing an additional and arbitrary constraint of unit variance, i.e.

$$\int_{-\infty}^{+\infty} K(y)\, y^2\, dy = 1 \tag{5}$$

He then uses an approximation to the integrated mean square error between $\hat{f}_n(x)$ and f(x), and he finds the optimizing K. The constraint (5) makes his solution suboptimal, because any moment constraint of the type

$$\int_{-\infty}^{+\infty} K(y)\, y^{2m} dy = 1 \quad , \quad m = 1, 2, \ldots \tag{5´}$$

could be equally justified and provide a different optimal K_m. Also an approximation to the integrated mean square error is a weak criterion. We will show how we can avoid the above disadvantages and still obtain the simple solution given in the abstract.

OPTIMAL KERNEL FUNCTION

In a series of papers, G. Wahba [4] - [5] recently compared the convergence properties of the following density estimators:

a) Ordinary Histogram
b) Certain orthogonal series estimates
c) Kernel type estimates
d) Polynomial and spline interpolation methods

Her result is an upper bound to the pointwise mean square error, expressed as

$$E\,[\hat{f}_n(x) - f(x)]^2 \leqslant D_j n^{-s(m,p)}$$

where D_j, j=a, b, c, d are constants corresponding to each of the above estimates, and

$$s(m,\, p) = (2m - 2/p) \,/\, (2m + 1 - 2/p)$$

where the true density f is assumed to belong to the class:

$$W_p^{(m)}(u) = \{f;\ f^{(i)} = \text{abs. cont.},\ i=1,..,m-1,\ \|f^{(m)}\|_p \leqslant u\}$$

where $\|f^{(m)}\|_p = \left|\int_s [f^{(m)}(t)]^p\, dt\right|^{1/p}$, $1 \leqslant p \leqslant \infty$ and $m \geqslant 2$, s=interval of definition of f.
Farrell [6] has shown that for any estimate $\hat{f}_n(x)$ and $f \varepsilon W_p^{(m)}(u)$, the pointwise mean square error cannot have a rate of convergence better than $n^{-s\,(m,p)}$

We will concentrate only on the kernel type density estimates. A uniform upper bound V to the class of densities is assumed known, i.e., $\sup_x |f(x)| \leqslant V$. We also assume $f \varepsilon W_p^{(m)}(u)$.

Let Q(y) be a real valued kernel function on $(-\infty, +\infty)$, satisfying the following conditions

$$\text{i)}\ \sup_{-\infty<y<+\infty} |Q(y)| < \infty\ ,\qquad \text{ii)}\ \int_{-\infty}^{+\infty} |Q(y)|\,dy < \infty\ ,$$

$$\text{iii)}\ \lim_{|y|\to\infty} |yQ(y)| = 0,\qquad \text{iv)}\ \int_{-\infty}^{+\infty} Q(y)\, dy = 1 \tag{6}$$

$$\text{v)}\ \int_{-\infty}^{+\infty} y^s Q(y)dt = 0\ \text{ for } s=1,..,m-1,\qquad \text{vi)}\ \int_{-\infty}^{+\infty} |y|^{m-1/p}\, |Q(y)|\,dy < \infty$$

We consider now the kernel estimate $\hat{f}_n(y) = (nh_n)^{-1} \sum_{k=1}^{n} Q[h_n^{-1}(y-x_k)]$

where $h_n \to 0$, $nh_n \to +\infty$ as $n \to +\infty$

We also define constants

$$B = V \cdot \int_{-\infty}^{+\infty} Q^2(y)\, dy\ ,\quad a = m - p^{-1}\ ,\quad q = (1 - p^{-1})^{-1}$$

$$A = \left[\int_{\infty}^{\infty} |Q(y)|\,|y|^{a}\,dy\right]^{2} \cdot [(m-1)!]^{2} \cdot [(m-1)\,q + 1]^{-2/q} \tag{7}$$

Then, the following upper bound was given by Wahba:

$$E[\hat{f}_n(y) - f(y)]^2 \leq V^2 A\, h_n^{2a} + B(nh_n)^{-1} \tag{8}$$

If we choose h_n in a way that will minimize the given bound, its resulting value is:

$$E[\hat{f}_n(y) - f(y)]^2 \leq C\cdot G(Q)\, n^{-s(m,q)}\, [1+0(1)] \tag{9}$$

where $0(1) \rightarrow 0$ as $n \rightarrow \infty$,

$$C = (1+2a)\,(2a)^{-2a}\, U^{2/(1+2a)}\, V^{2a/(1+2a)}$$

$$\cdot\, [(m-1)!\, [q(m-1) + 1]^{1/q}\,]^{-2/(1+2a)}$$

$$G(Q) = \int_{-\infty}^{+\infty} |Q(y)|\,|y|^{a} dy \cdot \left[\int_{-\infty}^{+\infty} Q^2(y)\, dy\right]^{a} \tag{10}$$

We let m=2 and we constrain Q(y) to be nonnegative and even around zero, having a finite support [-T,T]. Then, all constraints (a) are satisfied. Let $b = (2T)^{-1}$, $K(z) = Q(zT)$, $z\varepsilon[-1,1]$. Then, minimization of G(Q) is equivalent to the minimization of

$$H(K) = \int_{o}^{1} K(z)\, z^{a}\, dz \left[\int_{o}^{1} K^2(z)\, dz\right]^{a} = T^{-2a-1}\quad G(K) \tag{11}$$

under the constraints $o \leq K(z) \leq V_o$, $z\varepsilon$ [o,1] and $\int_{}^{1} K(z)\, dz = b$

To this end, we fix $\int_{o}^{1} K^2(z)dz = q$, and we seek the minimum of H(K) under a fixed q. Then, by varying $b^2 \leq q \leq bU_o$, we find the overall minimum. Hence, in the new problem we want to minimize

$\int_{o}^{1} K(z)\, z^{a} dz$, under the constraints:

$$o \leq K(z) \leq V_o, \quad \int_{o}^{1} K(z)dz = b, \quad \int_{o}^{1} K^2(z)dz = q \tag{12}$$

We observe that the optimal K(z) must be monotime decreasing on [o,1], because z^a is monotone increasing. Hence, the optimal K must have one of two forms

A) K(z) = positive for all $z\varepsilon[0,1]$
B) K(z) = positive for all $z\varepsilon[0,u]$ and zero for $z\varepsilon[u,1]$, $o<u<1$.

We use Lagrange multipliers A_1, B_1 to identify the optimal K. Let

$$Lu\,(K,A_1,\, B_1) = \int_{o}^{u} K(z)\, z^{a} dz - A_1 \int_{o}^{u} K(z)dz + 2^{-1}B_1 \int_{o}^{u} K^2(z)dz$$

$$= 2^{-1}B_1 \int_{o}^{u} [K(z) - B_1^{-1}(A_1 - z^{a})]^2 dz + D_1 \tag{13}$$

D_1 = independent of K(z)

A small investigation results in the conclusion that K minimizes L_u and is monotone increasing only for $A_1>0$, $B_1>0$. Its optimal value is:

$$K_o(z) = B_1^{-1}(A_1 - z^a), \quad z\varepsilon[o,u]$$

For solutions of type (A), we must have $A_1 \geqslant 1$, $u=1$, $B_1>0$. (14)

As a result, the region of q where (14) are satisfied, is

$$1 \leqslant qb^{-2} \leqslant 1 + (1 + 2a)^{-1} \tag{15}$$

For solutions of type (B), we must have

$$A_1 = u^a, \quad o<u<1, \quad B_1 > o \tag{16}$$

The corresponding region of q where (16) are satisfied, is

$$qb^{-2} \geqslant 1 + (1 + 2a)^{-1} \tag{17}$$

Let $H_o(q)$ be the optimal value of $H(K)$ for a fixed q. Then, using the above optimal K_o, we find the following expression for $H_o(q)$:

$$b^{-1-2a}\,H_o(q) = \begin{cases} (1+a)^{-1}[1-a(1+2a)^{-1/2}(qb^{-2}-1)^{1/2}](qb^{-2})^a \\ \qquad \text{for solutions (A) and region } 1\leqslant qb^{-2}\leqslant 1+(1+2a)^{-1} \\ (1+2a)^{-1}[1+(1+2a)^{-1}]^a \\ \qquad \text{for solutions (B) and region } qb^{-2}\geqslant 1+(1+2a)^{-1} \end{cases} \tag{18}$$

The optimal q for solutions (A) is achieved for $qb^{-2} = 1+(1+2a)^{-1}$. Any value of q in the region $qb^{-2} \geqslant 1+(1+2a)^{-1}$ achieves the same $H_o(q)$, but on intuitive grounds any solution with u=1 achieves better data smoothing than one with o<u<1. Hence, we choose $qb_o^{-2}=1+(1+2a)^{-1}$. The final expression for the corresponding optimal kernel function with support [-T, T] is:

$$Q(y) = (1+a^{-1})\,(2T)^{-1}\,[1-T^{-a}|y|^a] \tag{19}$$

for $|y|\leqslant T$, and $a = 2-p^{-1}$, $p\geqslant 1$.

REFERENCES

[1] Rosenblatt, M. "Remarks on some nonparametric estimates of a density function", Ann. Math. Statist., Vol. 27, pp. 832-837, 1956.

[2] Parzen, E., "On estimation of a probability density function and mode", Ann. Math. Statist., Vol. 33, pp. 1065-1076, 1962.

[3] Epanechnikov, V. A., "Nonparametric Estimates of a multivariate probability density", Theor. Prob. Applic., Vol. 14, pp. 153-158, 1969.

[4] Wahba, G., "Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation", Ann. Statist., Vol. 3, pp. 15-29, 1975.

[5] Wahba, G., "Interpolating spline methods for density estimation, I. Equi-spaced knots", Ann. Statist., Vol. 3, pp. 30-48, 1975.

[6] Farrell, R. H., "On best obtainable asymptotic rates of convergence in estimation of a density function at a point", Ann. Math. Statist., Vol. 43, pp. 170-180, 1972.

RANGE OF VALIDITY OF THE SAMPLING THEOREM

BALAJI SWAMINATHAN AND R. J. LEAKE
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana 46556

SUMMARY

The sampling theorem of Whittaker is widely abused and misquoted. As an illustration of this, a textbook [1] is paraphrased below.

A function x(t) such that its Fourier transform X(jω) equals zero for $|\omega| > \Omega$ is completely determined by its sample values $x(n\pi/\Omega)$. x(t) has the following cardinal series expansion.

$$x(t) = \sum_{n=-\infty}^{\infty} x\left(\frac{n\pi}{\Omega}\right) \frac{\sin(\Omega t-n\pi)}{\Omega t - n\pi} \qquad (1)$$

a. Let the "sampling theorem" above hold for $x_1(t)$. Let a second function x(t) equal $x_1(t)$ except at the sampling points where it equals zero. Clearly, the sampling theorem does not hold for x(t) although it has the same Fourier transform as $x_1(t)$.

b. Consider $x(t) = \sin \Omega t$ with $X(j\omega) = [\delta(\omega-\Omega) - \delta(\omega+\Omega)]\pi/j$ and $x(j\omega) = 0$, $|\omega| > \Omega$. But $x(n\pi/\Omega) = \sin n\pi = 0$. Clearly, the sampling theorem as stated above does not hold in this case.

c. Let $x(t) = t^2$; $X(j\omega) = -2\pi\delta^{(2)}(\omega)$ and $X(j\omega) = 0$ for $\omega \geq 0$. But the series

$$\sum_{n=-\infty}^{\infty} n^2 \frac{\sin \pi(t-n)}{\pi(t-n)} \qquad (2)$$

diverges for all but interger values of t.

The first two difficulties may be circumvented by initially specifying x(t) as the inverse Fourier transform of a function with support contained in $(-\Omega,\Omega)$. With this change it can be shown [2] that (1) holds provided X(jω) is a measurable function with $\int_{-\Omega}^{\Omega} |X(j\omega)| d\omega < \infty$. In order to provide a framework for determining precisely for what class of signals (1) holds, X(jω) is considered a distribution of compact support. Such distributions can be classified in a hierarchy of vector spaces corresponding to the classifcation of periodic distributions as in [3]

$$\overline{T}_r(\Omega) \subset \overline{C}_\infty(\Omega) \subset \overline{A}(\Omega) \subset \overline{C}(\Omega) \subset \overline{L}_\infty(\Omega) \subset \overline{L}_2(\Omega) \subset \overline{L}_1(\Omega) \subset \overline{M}(\Omega) \subset \overline{P}(\Omega) \subset \overline{D}(\Omega)$$

It can be shown that (1) holds for $X(j\omega) \in \overline{L}(\Omega)$. It is seen from (2) that (1) does not hold for some $X(j\omega) \in \overline{D}(\Omega)$ though a paper [4] claims it does. Campbell [5] has shown that, with the introduction of a convergence factor in (1), a sampling series is obtained which is valid for $X(j\omega) \in \overline{D}(\Omega)$.

REFERENCES

[1] H. Freeman, "Discrete-Time Systems," John, Wiley and Sons, Inc., 1965.
[2] R. P. Boas, Jr., "Entire Functions," Academic Press, 1954, pp. 219-221.
[3] R. E. Edwards, "Fourier Series," Vol, II, Holt, Rinehart and Winston, 1967.
[4] E. Pfaffelhuber, "Sampling Series for Band Limited Generalized Functions," IEEE Trans. Inform. Theory, It-17, 6, pp. 650-655, 1971.
[5] L. L. Campbell, "Sampling Theorem for the Fourier Transform of a Distribution with Bounded Support," SIAM J. Appl. Math., Vol. 16, 1968, pp. 626-636.

SINGULAR PERTURBATION ANALYSIS APPROACH FOR SYSTEMS WITH HIGHLY COUPLED DYNAMICS

ANTHONY J. CALISE
Dynamics Research Corporation
Wilmington, Massachusetts 01887

ABSTRACT

This paper proposes a procedure for applying singular perturbation methods to separately analyze state dynamics even when they are highly coupled. The intent is to expand the family of problems to which these methods can be successfully applied. In particular several problems in flight mechanics are identified. Numerical results for the minimum time to climb problem are given where the procedure is used to separate altitude and flight path angle dynamics to produce a closed-form solution for lift.

INTRODUCTION

By now there have been numerous papers discussing and illustrating the application of singular perturbation methods to the solution of optimal control problems [1-8]. The principle advantage cited is that the computational burden is significantly reduced by approximating the solution of the original problem with the solution of an appropriately matched series of lower order problems. However, when dealing with nonlinear systems, even low order models present a considerable computational burden if they require the solution of a nonlinear two-point boundary-value problem (TPBVP). Real time applications are generally limited to situations where a solution in feedback form can be found.

References [5-8] discuss the application of singular perturbation methods to optimal control problems in aircraft flight dynamics. In [8] it is shown that singular perturbation methods can be used to extend existing reduced order modeling methods to include the faster dynamics often ignored in flight performance studies. Furthermore, if the dynamics can be ordered on separate time scales such that one state equation appears in each boundary layer, then a nonlinear feedback control solution can be obtained.

The intent of this paper is to suggest an approach for matching asymptotic expansions that would permit the separate analysis of state dynamics even when they are highly coupled. An informal proof is given for convergence in a restricted, but important class of problems. Several examples dealing with optimal aircraft flight that fall within the class of problems for which the method is proven applicable are discussed. Numerical results for minimum time aircraft climbs are given to show that accurate solutions are obtainable for a more general class of problems.

PRELIMINARIES

The class of problems to which the discussion here on convergence is restricted is given by the following optimal control problem definition:

$$\dot{x} = f(y, u) \tag{1}$$

$$\epsilon \dot{y} = g(y, u) \tag{2}$$

$$\text{Min}\left\{ J = \int_{o}^{t_f} L(y, u)\, dt \right\}, \; t_f \text{ free} \tag{3}$$

$$x(0) = 0 \qquad x(t_f) = x_f \tag{4}$$

$$y(0) = y_i \qquad y(t_f) = y_f \tag{5}$$

where x and y are scaler state variables and u is a scaler control variable . The range of y_i and y_f is also constrained as will be defined shortly.

For linear systems the parameter ϵ is usually identified with a small system parameter. For nonlinear systems it is generally not possible to make such an identification so that ϵ is artificially introduced as a parameter for expansion of the solution about $\epsilon = 0$. The value $\epsilon = 1$ is then used to construct the series solution for the original system of dynamics. Although much of the theory for convergence of singular expansions relies on the assumption that $\epsilon << 1.0$, this does not appear to be a requirement for the successful application of singular perturbation methods. It seems to be more a function of of the system dynamics, control variable limits and boundary conditions of interest. In any case let us turn our attention to the solution of the control problem in (1-5).

Outer Solution

The necessary conditions for optimality in the zero order outer solution are:

$$\dot{x}_0 = f_0(y_0), \qquad \dot{\lambda}_{x_0} = 0 \tag{6}$$

$$H_0 = \lambda_x f_0(y) + L_0(y) = 0 \tag{7}$$

$$y_0 = \arg \{\min_{y} H_0(\lambda_x, y)\} \tag{8}$$

along with the boundary conditions in (4). It is assumed that $g(y, u) = 0$, which results from letting $\epsilon \to 0$ in (2), yields a unique solution for

$u = h(y)$. This is used in f and L to construct the functions f_0 and L_0 in (6) and (7). Equations (7) and (8) now represent two constraints for the unknowns λ_x and y, which is assumed to produce a unique solution:

$$y_0 = \arg\{\min_y[L_0(y)/f_0(y)]\} \tag{9}$$

$$\lambda_{x_0} = -L_0(y_0)/f_0(y_0) \tag{10}$$

Thus the zero order outer solution results in stationary values for y_0 and $u_0 = h(y_0)$.

Inner Solution

The inner solution is used to match y_0 to the boundary conditions in (5). In the initial boundary layer the time scale $\tau = t/\epsilon$ is used. The zero order necessary conditions are:

$$\frac{d}{d\tau} y^o = g(y^o, u), \qquad \frac{d}{d\tau} \lambda^o_y = -\frac{\partial H^o}{\partial y} \tag{11}$$

$$H^o = \lambda_{x_0} f(y^o, u) + \lambda_y g(y^o, u) + L(y^o, u) = 0 \tag{12}$$

$$u^o = \arg\{\min_u H(\lambda_{x_0}, \lambda_y, y^o, u)\} \tag{13}$$

where $y^o(0) = y_i$. Equations (12) and (13) are assumed to result in unique solutions for λ^o_y and u^o.

$$u^o = \arg\{\min_u[(\lambda_{x_0} f + L)/g]\} \tag{14}$$

$$\lambda^o_y = -[\lambda_{x_0} f(y^o, u^o) + L(y^o, u^o)]/g(y^o, u^o) \tag{15}$$

Note that (14) is a nonlinear control solution in feedback form $u^o = N(y^o)$, where λ_{x_0} is computed using (10) and y^o is obtained from the simultaneous integration of (11). The dynamics of x are obtained from the first order inner expansion equation

$$\frac{d}{d\tau} x^1 = f(y^o, u^o) \tag{16}$$

Integration of (11) and (16) using (14) to define the control produces $x(\tau)$ and $y(\tau)$ for the corresponding value of ϵ.

In the terminal boundary layer the time scale $\sigma = (t_f - t)/\epsilon$ is used and the necessary conditions are similar to (11)-(13) with $y^o(0) = y_f$. Hence a backward time solution $x(\sigma)$ and $y(\sigma)$ can be generated in an exactly parallel manner.

The proceeding discussion has been specialized to the case where x does not appear in the right hand side of (1) and (2), and u is a scaler. This was done to simplify the discussion on convergence of the matching procedure in the next section. However, it is easy to show that the conclusion regarding the feedback form of u is valid for the case where f, g and L are functions of x as well and u is a vector. Examples 2 and 3 illustrate the case where u has several components. When f, g or L depend on x, then y_0 and u are dependent on x and a first order correction for λ_x is needed to produce an accurate solution for u as a function of x and y. This is illustrated by Example 4.

THE PROCEDURE FOR MATCHING

For the successful application of the solution in (14), it is necessary that $y^o(\tau) \to y_0$ and that the solution is stable about y_0, or that

$$\lim_{y \to y_0} u^o(y) = u_0$$

which would then yield $g(y_0, u_0) = 0$. However, this is not sufficient to produce an accurate result. Given $x(\tau)$, $x(\sigma)$ and $y(\tau)$, $y(\sigma)$ it is possible to portray the boundary layer results as shown in Figure 1. At this point the discussion is confined to the range of boundary conditions such that both y_i and y_f are either greater than or less than y_0. When x_f is large there is a region of overlap at $y = y_0$. For this case the boundary layer and optimal solutions are nearly identical. For x_f small the boundary layers intersect before reaching steady-state. The optimal solution for this case never reaches y_0. Note that asymptotic accuracy is more dependent on the value of x_f rather than the magnitude of ϵ.

"Patching" the Boundary Layers

The situation depicted by Figure 1 suggests than an additional constraint should be introduced to the matching procedure that insures $dy/dx = 0$ at the boundary layer intersection point. This can be achieved by constraining y in the zero order outer solution. If in the unconstrained optimal solution to the original two state problem $y^*(t) < y_{max}$, then introducing the constraint $y < y_{max}$ to the perturbed problem can be viewed as not altering the original problem statement. A logical first choice for y_{max} is the value of y corresponding to the intersection of the curves labeled I.B.L. and T.B.L. Using $y_0 = y_{max}$ in (10) produces a new value for λ_{x_0} ($\lambda^1_{x_0}$) which in turn leads to a control solution u^{o1} in (14) that is asymptotic to u^1_0 as $y \to y_{max}$, where now $g(u^1_0, y_{max}) = 0$. This results in a new set of boundary layer solutions for x and y that intersect at a lower value of y. This procedure can be iterated until a sufficiently small value of dy/dx is observed at the intersection point.

For the optimal control problem defined by (1-5) the above procedure converges exactly to the optimal solution. This can be seen by noting that the Hamiltonian in (12) and the control solution in (14) are the same as that for the original unperturbed problem. The only thing that differs is the method of establishing the correct value for λ_x, which is constant in the original problem. Hence it can be evaluated at any point along the trajectory. Note that the equation in (10) for computing λ_{x_0} is also obtained using $H = 0$ in the original problem and evaluating at the point in the trajectory where $g(y, u) = 0$. If we use $y_0 = y^n_{max}$ in (10), where y^n_{max} is the converged value of y_{max} after n iterations, then $g(u_1^n, y^n_{max})$ would indeed be zero due to the asymptotic property of the boundary layers and the additional matching constraint that $dy/dx = 0$ at the boundary layer intersection point. Hence the constrained boundary layer solution exactly equals the optimal solution.

The argument at this point is not complete since it is possible to construct boundary layer arcs using (14) and (10) with y_0 constrained to $y_{max} < y^n_{max}$. This would result in a patched solution for which $dy/dx = 0$ over a finite interval of time. During such intervals we have that $g(y, u) = 0$ and λ_y^o = constant, or that $\partial H^o/\partial y = 0$. This is equivalent to $\frac{\partial H_0}{\partial y} = 0$, for which we have assumed a unique solution $y = y_0$, or the unconstrained solution in (8). Hence, boundary layer arcs constrained to values of y less than y^n_{max} are non-optimal.

The preceding arguments can easily be extended to the case where u is a vector.

Example 1

It is perhaps best to cement the concepts presented in the preceding section with a simple example. Consider the following optimal control problem:

$$\dot{x} = y \tag{18}$$

$$\epsilon \dot{y} = u, \qquad \epsilon = 1.0 \tag{19}$$

$$\text{Min}\ \{J = \int_0^{t_f} 1 + u^2/2\ dt\} \tag{20}$$

$$x(0) = 0 \qquad x(t_f) = x_f \tag{21}$$

$$y(0) = 0 \qquad y(t_f) = 0 \tag{22}$$

It is easy to show from the necessary conditions that this problem admits to the following open-loop solution:

$$x/x_f = 3\,t_1^2 - 2\,t_1^3 \tag{23}$$

$$y/y_{max} = 4(t_1 - t_1^2) \tag{24}$$

$$u = \sqrt{2}\,(1-2t_1) \qquad (25)$$

where $t_1 \equiv t/t_f$ and

$$x_f = t_f^2/3\sqrt{2} \qquad (26)$$

$$y_{max} = t_f/2\sqrt{2} \qquad (27)$$

The plot of y versus x is given in Figure 2. Note that $dy/dx = 0$ occurs only at $x/x_f = .5$ independent of x_f. Hence we should obtain $y_0 = \infty$ in the perturbed solution and it is necessary to introduce an artificial constraint $y < y_{max}$ right at the outset.

The necessary conditions for the zero order solution are:

$$u = 0 \qquad (28)$$

$$\dot{x} = y\,, \qquad x(0) = 0,\ x(t_f) = x_f \qquad (29)$$

$$H_0 = \lambda_x y + 1 = 0 \qquad (30)$$

$$y_0 = \arg \underset{\substack{y \\ y < y_{max}}}{\{\min H_0\}} \qquad (31)$$

for which we obtain the constrained solution

$$y_0 = y_{max} \qquad (32)$$

$$\lambda_{x_0} = -1/y_{max} \qquad (33)$$

The zero order boundary layer necessary conditions are:

$$\dot{y} = u \qquad y(0) = 0 \qquad (34)$$

$$H^o = \lambda_{x_0}\, y + \lambda_y\, u + 1 + u^2/2 = 0 \qquad (35)$$

$$\frac{\partial H}{\partial u} = \lambda_y + u = 0 \qquad (36)$$

Using (36) to eliminate λ_y in (35), we have the following feedback solution for u

$$u = \pm \sqrt{2\,H_0(y)} \qquad (37)$$

where $H_0(y)$ is defined in (30). The plus sign is used for the initial boundary layer and the minus sign is used for the terminal boundary layer. Note that from (30) and (33), $u \to 0$ as $y \to y_{max}$. It remains to show that the solution in (37) produces the same result as that in (23) - (27).

Using (37) in (34) we have

$$\dot{y} = \sqrt{2(1-y_1)} \tag{38}$$

where $y_1 = y/y_{max}$. From the definition for t_1 we have

$$\frac{dy_1}{dt_1} = \frac{t_f}{y_{max}} \sqrt{2(1-y_1)} \tag{39}$$

where $y_{max} = t_f/2\sqrt{2}$. So we are interested in the solution to

$$\frac{dy_1}{dt_1} = 4\sqrt{(1-y_1)} \tag{40}$$

Defining $y_1 \equiv at_1 + bt_1^2$ and squaring both sides of (40) we have

$$(a + 2\,bt_1)^2 = 16(1-a\,t_1 - bt_1^2) \tag{41}$$

whereby equating like coefficients we obtain $a = -b = 4$. Hence, the initial boundary layer solution is

$$\begin{aligned} y(\tau) &= 4\,(t_1 - t_1^2)\,, & t_1 &< .5 \\ y(\tau) &= y_{max}\,, & t_1 &\geq .5 \end{aligned} \tag{42}$$

which agrees exactly with (24) for $t_1 < .5$.

APPLICATION

In this section three examples dealing with optimal aircraft flight are used to illustrate the usefulness of the constrained matching technique. In each case the original problem is actually a reduced order model of the complete six-degree-of-freedom equations of aircraft motion. However we are still left with a TPBVP to solve. Further model order reduction using singular perturbation theory leads to a control solution in feedback form.

The first two problems are examples where the aircraft equations of motion belong to the class defined by (1) and (2), but with several control variables. In the third problem there are three state equations, but it is possible to reduce the problem to the format given in (1-5) using the concept of multiple time scales. However, the righthand sides (f and g functions) are dependent on all of the state variables. The constrained matching technique is shown to apply equally well for this example when the solution is corrected to first order.

Example 2

Consider the problem of finding the minimum fuel control for flying to a specified range. In [9] the following equations are used to define the aircraft dynamics:

$$\dot{x} = V \tag{43}$$

$$\epsilon \dot{E} = (T-D)\, V/W\,, \qquad \epsilon = 1.0 \tag{44}$$

where the energy state E is the sum of kinetic plus potential energy per unit weight

$$E = h + V^2/2g \tag{45}$$

and x is down range position, W is weight, T is thrust, D is drag, h is altitude and V is velocity. These equations assume that small flight path angles pervail over most of the trajectory [10]. Under these assumptions T and D can be viewed as nonlinear functions of h, E and throttle setting (δ)

$$T = T(h,\ E,\ \delta) \tag{46}$$

$$D = D(h,\ E) \tag{47}$$

In general the fuel flow (σ) is dependent on h, E and δ, so

$$J = \int_0^{t_f} \sigma(h,\ E,\ \delta)\, dt \tag{48}$$

Hence the control variables for the original problem defined above are h and δ.

The necessary conditions for the zero order outer solution are:

$$T = D \tag{49}$$

$$H_0 = \lambda_x V + \sigma = 0 \tag{50}$$

$$h_0,\ E_0 = \arg\{\underset{\substack{h,\,E \\ T\,=\,D}}{\min}(\sigma/V)\} \tag{51}$$

which establishes the optimum h and E (or equivalently h and V) steady-state cruise condition†. The zero order inner solution is used to establish the optimum climb and descent schedule $h^o(E)$. The necessary conditions are:

$$H^o = \lambda_{x_0} V + \lambda_E (T-D)\, V/W + \sigma = 0 \tag{52}$$

---

† There have been several papers addressing the optimality of cruise [11, 12]. This yet unresolved issue is beyond the scope of this paper.

$$h^\circ, \delta^\circ = \arg\{\min_{h^\circ, \delta^\circ} [\frac{H_0(V)}{(T-D)V/W}]\} \qquad (53)$$

where from (50)

$$\lambda_{x_0} = -\frac{\sigma(h_0, V_0, \delta_0)}{V_0} \qquad (54)$$

There are two solution paths obtained from (53), one for $\dot{E} > 0$ and one for $\dot{E} < 0$.

Examples of minimum fuel trajectories taken from [9] are shown in Figure 3. The long range solutions ($R > 457$ km) reach a cruse condition at $h = 9.8$ km and $V = 225$ m/sec. The climb and descent paths are always the same and the range differences are accounted for at the cruise point. For short range missions ($R < 457$ km) there is no cruise phase and the climb and descent arcs match at different altitude and air speed points which can be obtained by carrying out a constrained optimization ($h < h_{max}$) in (51). The resulting inner solution obtained from (53) would exactly match the climb and descent profiles of Figure 3. In this case E would be plotted versus x with the I. B. L. starting at $x = 0$ and the T. B. L. starting at x = final range. Altitude is constrained using the iterative matching procedure until $dE/dx = 0$ at the intersection point.

Example 3

There are other important problems dealing with aircraft flight that fall within the class of problems defined by (1-5). For example the problem of minimum time turns in three-dimensions is treated using perturbation methods in [7]. The key result is depicted in Figure 4. For large heading changes the boundary layer paths 1 ($\dot{E} < 0$) and 2 ($\dot{E} > 0$) are asymptotic to the h and V point (point B) where the aircraft achieves its highest turn rate at zero $\dot{E}$. The analysis in [7] can be extended to the case of small heading changes merely by finding new maximum sustained ($\dot{E} = 0$) turn rate points (B') subject to the constraint $h > h_{min}$. This results in new boundary layer paths as shown in Figure 5. The optimum B' is achieved when the amount of turning along paths 1 and 2 equals the specified heading change with zero time spent at B'.

Example 4

The problem statement for minimum time aircraft climbs is given as:

$$\dot{E} = (T-D)\,V/W \qquad (55)$$

$$\epsilon\,\dot{h} = V \sin\gamma, \qquad \epsilon = 1 \qquad (56)$$

$$\epsilon^2\,\dot{\gamma} = g\,[L/W - \cos\gamma]/V \qquad (57)$$

$$\min\{J = \int_0^{t_f} dt\} \qquad (58)$$

$$E(0) = E_i \qquad E(t_f) = E_f$$

$$h(0) = h_i \qquad h(t_f) = h_f$$

$$\gamma(0) = \gamma_i \qquad \gamma(t_f) = \gamma_f \tag{59}$$

where the previously undefined variables are lift (L) and flight path angle (γ). Note from the placement of ϵ that each of the equations are ordered on separate time scales [8]. The zero order outer necessary conditions are:

$$L = W, \quad \gamma = 0 \tag{60}$$

$$H_0 = \lambda_E(T-D)V/W + 1 = 0 \tag{61}$$

$$h_0 = \arg\{\max_h \dot{E}(h, E)\} \tag{62}$$

which produces an optimum climb schedule $h_0 = f_h(E)$ as shown in Figure 4, [10]. We are concerned here with the transitional arcs from γ_i, h_i to $\gamma = 0$, $h \Rightarrow h_0$. In [13, 14] h and γ are treated as fast variables that vary on the same time scale (ϵ to the first power in (57)). It is necessary to solve a TPBVP in the boundary layer. Hence the zero order inner solutions in [13] can be viewed as exact optimal solutions for constant E transitions (to zero order E is constant in the boundary layers). This example is confined to comparing the separate boundary layer solutions for h and γ to the optimal solutions in [13].

In [15] the boundary layer dynamics for the problem statement in (55-59) are analyzed to zero and first order. It is shown that the analysis of h and γ dynamics in separate boundary layers leads to a nonlinear feedback control solution. The control solution was used to define L and the original system of equations (56, 57) were integrated.

Numerical results for the same aircraft used in [13] are given in Figure 6 for h(0) = 40000 feet, V(0) = 484 ft/sec and $\gamma(0) = -2$ deg. The energy level for h(0), V(0) is E(0) = 43637.5 ft and the corresponding optimum altitude for maximum $\dot{E}$ computed using (62) is $h_0 = 7500$ feet. The E state was held constant in the simulation for comparison to the optimal solutions in [13]. The curves labeled A correspond to a forward and backward time integration of (56) and (57) for $\gamma_{max} = \pi/2$. Note that the boundary layers intersect before reaching their steady state values. This is a consequence of the higly coupled h and γ dynamics.

The curves labeled B correspond to the constrained matched solution after four iterations. Also shown is the numerical result for the two-state solution given in [13]. Figure 7 presents the same comparison for $\gamma(0) = 60^\circ$. Convergence in this case was obtained after three iterations. In both cases the minimum values of γ agree to within a degree for both solution methods.

Figures 8 and 9 compare the first order solution in [15] with the optimal solution in [13]. Note that a remarkably accurate result is obtained.

SUMMARY

A state constraint approach has been outlined for obtaining matched asymptotic solutions for systems of equations with highly coupled dynamics. The method extends the range of problems to which singular perturbation methods can be applied and, more importantly, leads to control solutions in feedback form. The method is exact for the case of two state equations where the derivatives and the performance index are independent of one of the state variables. However, the approach seems to be applicable to the case where the derivatives and the performance index are functions of both state variables. It also is applicable to higher order systems when used in conjunction with the multiple time scales approach.

REFERENCES

[1]. P. V. Kokotovic, R. E. O'Malley, Jr., and P. Sannuti, "Singular Perturbations and Order Reduction in Control Theory-An Overview," presented at the 6th World IFAC Congr., Aug. 1975, paper 51:3.

[2]. P. V. Kokotovic, "A Control Engineer's Introduction to Singular Perturbations", in Singular Perturbations: Order Reduction in Control System Design. New York, ASME, 1972.

[3]. Wilde, R. R., Kokotovic, P. V., "Optimal Open-and Closed-Loop Control of Singularly Perturbed Linear Systems," IEEE Trans. on A. C., vol. AC-18, no. 6, December 1973.

[4]. Reddy, P. B., Sannuti, P., "Optimal Control of a Coupled-Core Nuclear Reactor by a Singular Perturbation Method," IEEE Trans. on A.C., Vol. AC-20, No. 6, December, 1975.

[5]. Kelley, H. J., "Flight Path Optimization with Multiple Time Scales," J. Aircraft, vol. 8, no. 4, April 1971.

[6]. Kelley, H. J., and Edelbaum, T. N.,"Energy Climbs, Energy Turns, and Asymptotic Expansions" J. Aircraft, Vol. 7, Jan. 1970.

[7]. Calise, A. J., "Singular Perturbation Methods for Variational Problems in Aircraft Flight", IEEE Trans. on A.C., vol. AC-21, no. 3, June 1976.

[8]. Calise, A. J., "Extended Energy Management for Flight Performance Optimization," presented at the 13th AIAA Aerospace Sciences Meeting, Jan. 20-220, 1975, paper 75-30.

[9]. Ertzberger, H., Barman, J. F., McLean, J. D., "Optimum Flight Profiles for Short Haul Missions", AIAA Guidance & Control Conference, Boston, Mass., August 1975.

[10]. Bryson, A. E., Jr., Desai, M. N., and Hoffman, W. C., "Energy-State Approximation in Performance Optimization of Supersonic Aircraft," Journal of Aircraft, vol. 6, Nov. - Dec. 1969.

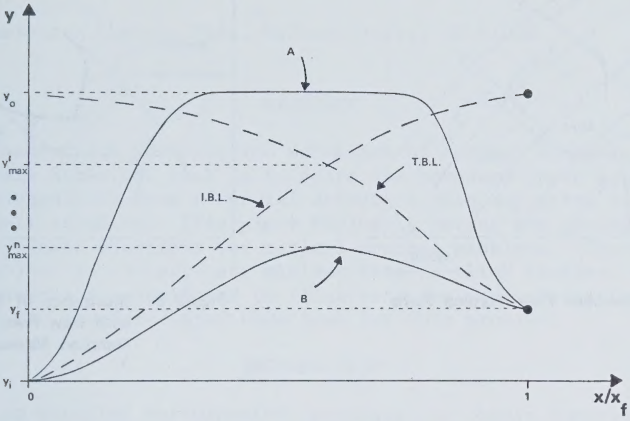
[11]. Speyer, J. L., "On the Fuel Optimality of Cruise", Journal of Aircraft, vol. 10, no. 12, December 1973.

[12]. Schultz, R., "Fuel Optimality of Cruise", Journal of Aircraft vol. 11, no. 9, September, 1974.

[13]. Ardema, M. D., "Solution of the Minimum Time-To-Climb Problem by Matched Asymptotic Expansions", AIAA Journal, vol. 14, no. 7, July 1976.

[14]. Ardema, M. D., "Singular Perturbations In Flight Mechanics", NASA TM-X-62-380, August 1974, revised October, 1975.

[15]. Calise, A. J., "Extended Energy Management for Flight Performance Optimization", revision submitted for publication in the AIAA Journal.



CURVE A - Boundary Layer and Optimal Solutions for Large x_f

CURVE B - Optimal Solution for Small x_f

CURVE I.B.L. - Initial Boundary Layer Solution for Small x_f

CURVE T.B.L.- Terminal Boundary Layer Solution for Small x_f

Figure 1. Comparison of Optimal and Boundary Layer Solutions

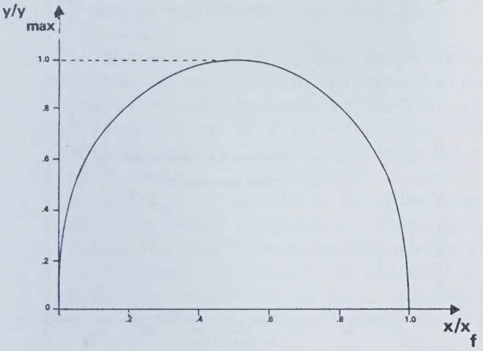


Figure 2. Optimal Solution for x and y in Example 1.

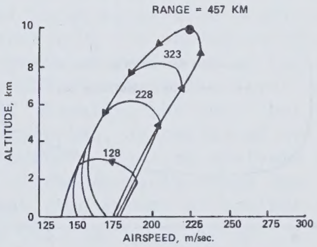


Figure 3. Speed-Altitude Profiles, CTOL Aircraft

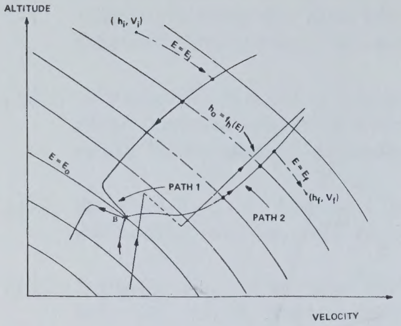

ALTITUDE
(h_i, V_i)
E = E_i
h_o = f_h(E)
E = E_o
PATH 1
B
E = E_f
(h_f, V_f)
PATH 2
VELOCITY

Figure 4. Minimum Time Turning Paths

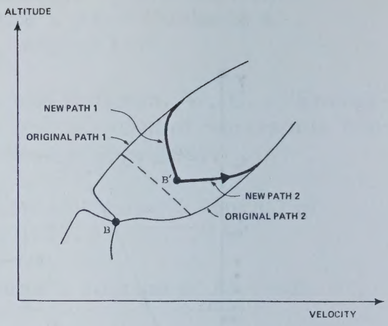

ALTITUDE
NEW PATH 1
ORIGINAL PATH 1
B'
NEW PATH 2
ORIGINAL PATH 2
B
VELOCITY

Figure 5. Illustration of New Paths 1 and 2 with New Point B′ for a Short Duration Maneuver

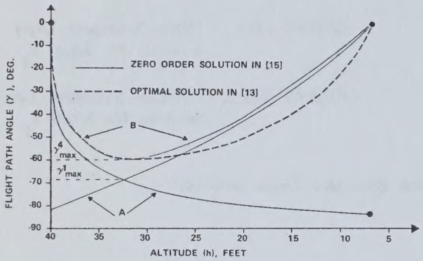


Figure 6. Comparison of the Zero Order Separation of h and γ with the Two–State Zero Order Solution for γ (0) = –2.0

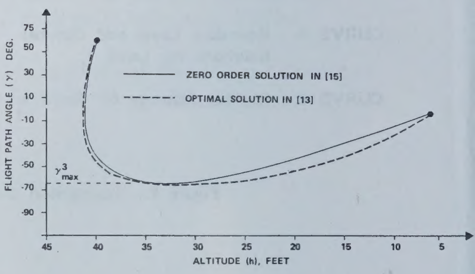


Figure 7. Comparison of the Zero Order Separation of h and γ with the Two–State Zero Order Solution for γ (0) = 60.

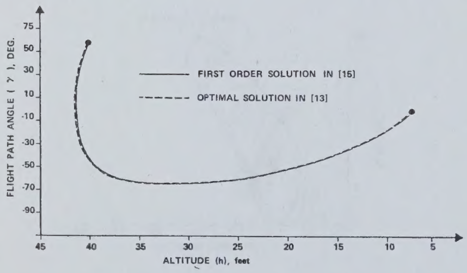


Figure 8. Comparison of the First Order Separation of h and γ with the Two–State Zero Order Solution for γ (0) = –2.0

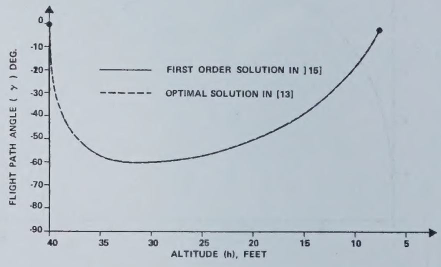


Figure 9. Comparison of the First Order Separation of h and γ with the Two–State Zero Order Solution for γ(0) = 60.

CHARACTERISTICS OF THE BOUNDARY-LAYER EQUATIONS OF THE MINIMUM TIME-TO-CLIMB PROBLEM

MARK D. ARDEMA
Ames Research Center, NASA, Moffett Field, CA 94035

ABSTRACT

In many singular perturbation solutions of optimal control problems, the most difficult numerical task is to solve the boundary-layer equations. However, these equations have a special structure that may often be used to expedite their solution. This paper begins by noting the general nature of the boundary-layer equations for optimal control problems. These results are then applied to the aircraft minimum time-to-climb problem. A specific numerical example is considered to illustrate the characteristics of the solution of the boundary-layer equations for this problem.

INTRODUCTION

In using singular perturbation techniques to obtain numerical solutions of nonlinear optimal control problems, the most difficult and costly part of the computation often is to determine suitable solutions of the zeroth-order boundary-layer equations (ZOBLE's). The purpose of these equations is to model the nonlinear dynamical behavior of the "fast" variables. (The higher order equations are linear and thus, in principle, pose no computational problems.) If there are m fast state variables, then the ZOBLE's are of order 2m. The zeroth-order outer solution evaluated at the boundary is a stationary point of these equations and a solution must be found which decays exponentially to this stationary point and satisfies m boundary conditions. Such a solution is called here a boundary-layer solution.

It would, of course, be helpful, before attempting a numerical solution, if one knew whether boundary-layer solutions exist. This is one of the aims of singular perturbation theory. Unfortunately, applications of singular perturbation theory to *nonlinear* optimal control problems have received relatively little attention. Two notable contributions are those of Hadlock [1] and Freedman and Kaplan [2].

Not surprisingly, the known results give two essential requirements (in addition to some technical restrictions) that are sufficient for the existence of boundary-layer solutions. The first is an eigenvalue condition ensuring a form of local stationary-point stability; the second is a restriction that the initial conditions be in the domain of influence of the stationary point. In practice, it is often possible to check the first condition but rarely possible to check the second. Usually, the best that can be done is a local analysis of the stability properties, accomplished by linearizing the equations about the outer solution. Such an analysis gives insight into the nature of the ZOBLE's and, if the analysis shows that the eigenvalue condition is met, inspires at least some amount of confidence that a boundary-layer solution can be found.

Aiken and Lapidus [3] have suggested solving the linearized ZOBLE's as an approximation of the nonlinear equations. This is justified on the basis that the boundary-layer solution decays with a large exponential. This procedure would not generally seem to be justified, however, because the purpose of the ZOBLE's is, in fact, to model the nonlinear behavior of the fast variables on a time scale stretched in proportion to the rate of decay.

Nevertheless, the savings in computational effort with this approximation are substantial and this approach should be investigated in applications. One use of the linear solution is to provide a means of predicting the m unknown initial conditions for the nonlinear ZOBLE's.

In the next section, the local properties of the ZOBLE's of a general nonlinear optimal control problem are investigated. In the subsequent section, these results are applied to the minimum time-to-climb (MTC) problem. The linearized ZOBLE's are obtained and solved. The final section presents a numerical example that illustrates the results.

LOCAL ANALYSIS OF ZEROTH-ORDER BOUNDARY-LAYER EQUATIONS FOR A GENERAL OPTIMAL CONTROL PROBLEM

Consider the nonlinear dynamical system:

$$\left.\begin{aligned} \frac{d\underline{x}}{dt} &= \underline{f}(\underline{x},\underline{y},\underline{u}) \ ; \quad \underline{x}(0) = \underline{X} \\ \varepsilon \frac{d\underline{y}}{dt} &= \underline{g}(\underline{x},\underline{y},\underline{u}) \ ; \quad \underline{y}(0) = \underline{Y} \end{aligned}\right\} \tag{1}$$

where $\underline{x}$, $\underline{y}$, and $\underline{u}$ are n-, m-, and r-dimensional vectors, respectively, and ε is a parameter. We call $\underline{x}$ the slow state variable and $\underline{y}$, the fast state variable. It is desired to transfer the system from $t = 0$ to $t = T$ while minimizing

$$J = \int_o^T \phi(\underline{x},\underline{y},\underline{u})dt \tag{2}$$

where T is not prescribed. Several generalizations of this problem are possible (cf. [4]). We assume that a unique, unbounded, nonsingular, optimal control exists. Under well-known smoothness and admissibility assumptions, application of the maximum principle results in the following two-point boundary value problems for the optimal control and trajectory:

$$\left.\begin{aligned} \frac{d\underline{x}}{dt} &= \underline{f}(\underline{x},\underline{y},\underline{u}) \ ; \quad \underline{x}(0) = \underline{X} \\ \varepsilon \frac{d\underline{y}}{dt} &= \underline{g}(\underline{x},\underline{y},\underline{u}) \ ; \quad \underline{y}(0) = \underline{Y} \\ \frac{d\underline{\lambda_x}}{dt} &= -\underline{H_x}(\lambda_o,\underline{\lambda_x},\underline{\lambda_y}, \underline{x},\underline{y},\underline{u}) \ ; \quad \underline{\lambda_x}(T) = \underline{0} \\ \varepsilon \frac{d\underline{\lambda_y}}{dt} &= -\underline{H_y}(\lambda_o,\underline{\lambda_x},\underline{\lambda_y},\underline{x},\underline{y},\underline{u}) \ ; \quad \underline{\lambda_y}(T) = \underline{0} \\ \underline{0} &= \underline{H_u}(\lambda_o,\underline{\lambda_x},\underline{\lambda_y},\underline{x},\underline{y},\underline{u}) \end{aligned}\right\} \tag{3}$$

where

$$H = \lambda_o\phi + \underline{f}'\underline{\lambda_x} + \underline{g}'\underline{\lambda_y} = 0 \tag{4}$$

Column vectors are underlined (matrices will be double-underlined) and the prime denotes transpose. In the sequel, functional dependence will frequently be omitted for brevity when no confusion results.

The ZOBLE's for the initial boundary layer are obtained from (3) by transforming the independent variable to $\tau = t/\varepsilon$, setting $\varepsilon = 0$, and setting the slow variables to their initial values. The result is

$$\left.\begin{aligned} \frac{d\underline{y}_o^{il}}{d\tau} &= \underline{g}(\underline{X},\underline{y}_o^{il},\underline{u}_o^{il}) \; ; \quad \underline{y}_o^{il}(0) = \underline{Y} \\ \frac{d\underline{\lambda}_{y_o}^{il}}{d\tau} &= -\underline{H}_y(\lambda_o,\underline{\lambda}_{x_o}^{o}(0),\underline{\lambda}_{y_o}^{il},\underline{X},\underline{y}_o^{il},\underline{u}_o^{il}) \\ \underline{0} &= \underline{H}_u(\lambda_o,\underline{\lambda}_{x_o}^{o}(0),\underline{\lambda}_{y_o}^{il},\underline{X},\underline{y}_o^{il},\underline{u}_o^{il}) \end{aligned}\right\} \quad (5)$$

where, e.g., y_o^{il} denotes a zeroth-order initial boundary-layer variable and $\lambda_{x_o}^{o}$, a zeroth-order outer variable. Since we are interested in the behavior of these equations near the zeroth-order outer solution evaluated at $t = 0$, the dependent variables are linearized according to

$$\underline{y}_o^{il}(\tau) = \underline{y}_o^{o}(0) + \underline{\alpha}(\tau) \; ; \quad \underline{\lambda}_{y_o}^{il}(\tau) = \underline{\lambda}_{y_o}^{o}(0) + \underline{\beta}(\tau) \; ; \quad \underline{u}_o^{il}(\tau) = \underline{u}_o^{o}(0) + \underline{\gamma}(\tau) \quad (6)$$

Substituting in (5), expanding to first order, and recalling that the outer solution is a stationary point results in

$$\left.\begin{aligned} \frac{d\underline{\alpha}}{d\tau} &= \underline{\underline{g}}_y\underline{\alpha} + \underline{\underline{g}}_u\underline{\gamma} \\ \frac{d\underline{\beta}}{d\tau} &= -\underline{\underline{H}}_{yy}\underline{\alpha} - \underline{\underline{H}}_{y\lambda_y}\underline{\beta} - \underline{\underline{H}}_{yu}\underline{\gamma} \\ \underline{0} &= \underline{\underline{H}}_{uy}\underline{\alpha} + \underline{\underline{H}}_{u\lambda_y}\underline{\beta} + \underline{\underline{H}}_{uu}\underline{\gamma} \end{aligned}\right\} \quad (7)$$

It is understood that all matrices in these equations are evaluated on the outer solution at $t = 0$. Since by assumption $\underline{\underline{H}}_{uu}$ is negative definite, it is invertible and the last of (7) gives

$$\underline{\gamma} = -\underline{\underline{H}}_{uu}^{-1}(\underline{\underline{H}}_{uy}\underline{\alpha} + \underline{\underline{H}}_{u\lambda_y}\underline{\beta}) \quad (8)$$

Also, from (4),

$$\underline{\underline{H}}_{y\lambda_y} = \underline{\underline{g}}_y' \; ; \quad \underline{\underline{H}}_{u\lambda_y} = \underline{\underline{g}}_u' \quad (9)$$

so that the first two in (7) become

$$\left.\begin{aligned} \frac{d\underline{\alpha}}{d\tau} &= \underline{\underline{A}}\,\underline{\alpha} + \underline{\underline{B}}\,\underline{\beta} \\ \frac{d\underline{\beta}}{d\tau} &= \underline{\underline{C}}\,\underline{\alpha} - \underline{\underline{A}}'\,\underline{\beta} \end{aligned}\right\} \quad (10)$$

where

$$\left.\begin{aligned} \underline{\underline{A}} &= \underline{\underline{g}}_y - \underline{\underline{g}}_u\underline{\underline{H}}_{uu}^{-1}\underline{\underline{H}}_{uy} \\ \underline{\underline{B}} &= -\underline{\underline{g}}_u\underline{\underline{H}}_{uu}^{-1}\underline{\underline{g}}_u' \\ \underline{\underline{C}} &= -\underline{\underline{H}}_{yy} + \underline{\underline{H}}_{yu}\underline{\underline{H}}_{uu}^{-1}\underline{\underline{H}}_{uy} \end{aligned}\right\} \quad (11)$$

Note that $\underline{\underline{B}}$ and $\underline{\underline{C}}$ are symmetric and that $\underline{\underline{B}}$ is positive definite. If the linearized equations (10) are used to approximate the nonlinear equations (5), the initial conditions are $\underline{\alpha}(0) = \underline{Y} - \underline{y}_o^o(0)$.

To construct boundary-layer solutions, (10) must have at least m stable modes. However, to ensure that the terminal ZOBLE's also have boundary-layer solutions, (10) must also have at least m stable modes when integrated backward with the coefficients evaluated at $t = T$. This leads to the well-known condition that the $2m \times 2m$ matrix

$$\underline{\underline{G}} = \begin{bmatrix} \underline{\underline{A}} & \underline{\underline{B}} \\ \underline{\underline{C}} & -\underline{\underline{A}}' \end{bmatrix} \tag{12}$$

must have m eigenvalues with positive and m with negative real parts for all $0 \leq t \leq T$. The form of the matrix $\underline{\underline{G}}$ is familiar from singular perturbation theory for linear systems with quadratic performance criteria (cf. [4] and [5]). The only differences are that for linear/quadratic problems $\underline{\underline{H_{uy}}} = \underline{\underline{0}}$ and $\underline{\underline{H_{yy}}}$ is assumed to be negative semidefinite, and thus $\underline{\underline{C}}$ is positive semidefinite.

The eigenvalue requirement is not as restrictive as it first appears. O'Malley and Fung [5] observed that for the linear/quadratic problem, $-S$ is an eigenvalue of $\underline{\underline{G}}$ whenever S is. Since this result depends only on the symmetry of $\underline{\underline{B}}$ and $\underline{\underline{C}}$, it applies also to the $\underline{\underline{G}}$ for the nonlinear problem. Thus the eigenvalue requirement may be stated as "all eigenvalues of $\underline{\underline{G}}$ must have nonzero real parts throughout $0 \leq t \leq T$." Another property of $\underline{\underline{G}}$ for linear/quadratic systems, i.e., that $\underline{\underline{G}}$ has no purely imaginary eigenvalues, depends on the positive semidefiniteness of $\underline{\underline{C}}$ and thus does not hold for $\underline{\underline{G}}$ in the nonlinear system.

Since the coefficients of the characteristic equation for (12) are real, not only is $-S$ a root if S is but $\bar{S}$ (bar denotes complex conjugate) and thus $-\bar{S}$ are also:

$$S \text{ a root} \Rightarrow -S,\ \bar{S},\ -\bar{S} \text{ roots} \tag{13}$$

This relation is useful for the analysis of specific problems.

The pattern of the roots as given by (13) causes numerical difficulties in determining boundary-layer solutions. System (5) may be viewed for computational purposes as a two-point boundary value problem with the terminal conditions imposed at some sufficiently large but arbitrary time. The presence of unstable modes generally will cause convergence problems for shooting methods since it will be impossible to completely eliminate the unstable modes in a numerical solution, causing the solutions to eventually grow. Further difficulties are created by the need to integrate functions of boundary-layer variables on the interval $[0,\infty)$ for use in the first-order equations. Thus numerical solution of nonlinear ZOBLE's is somewhat of an art. One possible approach is to derive the optimal control problem for which (5) are the state and adjoint equations. This was done by Kelley [6] for scalar $\underline{x}$ and $\underline{y}$. The method of steepest descent could then be applied.

For low-order systems, (13) leads to some general observations. For example, if there is one fast state variable, (13) implies that the two roots of $\underline{\underline{G}}$ will either be (i) purely real and equal in absolute value with opposite signs, (ii) purely imaginary and equal in absolute value with opposite signs, or (iii) both zero. Case (i) is the one required by the eigenvalue condition. Kelley [6] showed that this is equivalent to the strengthened form of the Legendre-Clebsch condition for the reduced problem for this case. For $m = 2$, if one eigenvalue $S = a + ib$ of $\underline{\underline{G}}$ has been found, then:

$$\begin{aligned} &\text{(i) } a \neq 0,\ b \neq 0 \Rightarrow \text{eigenvalue condition satisfied} \\ &\text{(ii) } a = 0 \Rightarrow \text{eigenvalue condition not satisfied} \\ &\text{(iii) } a \neq 0,\ b = 0 \Rightarrow \text{no information} \end{aligned} \tag{14}$$

MINIMUM TIME-TO-CLIMB PROBLEM

We begin by summarizing the local analysis of the ZOBLE's of the MTC problem as given in [4] and [7]. These references should be consulted for the nomenclature and algebraic details. The initial ZOBLE's are

$$\left.\begin{aligned} \frac{dh_o^{i1}}{d\tau} &= v_o^{i1} \sin \gamma_o^{i1} \\ \frac{d\gamma_o^{i1}}{d\tau} &= \frac{1}{v_o^{i1}} (L_o^{i1} - \cos \gamma_o^{i1}) \\ \frac{d\lambda_{h_o}^{i1}}{d\tau} &= \frac{\lambda_{h_o}^{i1}}{v_o^{i1}} \sin \gamma_o^{i1} - \lambda_{E_o} \frac{\partial p_o^{i1}}{\partial h_o^{i1}} - \frac{\lambda_{\gamma_o}^{i1}}{{v_o^{i1}}^3} (L_o^{i1} - \cos \gamma_o^{i1}) \\ \frac{d\lambda_{\gamma_o}^{i1}}{d\tau} &= -\lambda_{h_o}^{i1} v_o^{i1} \cos \gamma_o^{i1} - \frac{1}{v_o^{i1}} \lambda_{\gamma_o}^{i1} \sin \gamma_o^{i1} \end{aligned}\right\} \tag{15}$$

where the control law is

$$L_o^{i1} = \frac{\lambda_{\gamma_o}^{i1}}{2 {v_o^{i1}}^2 B_o^{i1} w \lambda_{E_o}^o(0)} \tag{16}$$

The boundary conditions on (15) are either

$$h_o^{i1}(0) = h_o \text{ and } \gamma_o^{i1}(0) = \gamma_o \quad \text{(problem A)} \tag{17}$$

$$h_o^{i1}(0) = h_o \text{ and } \lambda_{\gamma_o}^{i1}(0) = 0 \quad \text{(problem B)} \tag{18}$$

To linearize (15) about the zeroth-order outer solution evaluated at $t = 0$, recall that $\gamma_o^o(0) = 0$, $\lambda_{h_o}^o(0) = 0$, and $L_o^o(0) = 1$ and set

$$\left.\begin{aligned} &h_o^{i1}(\tau) = h_o^o(0) + \alpha(\tau) \ ; \quad \gamma_o^{i1}(\tau) = \beta(\tau) \\ &\lambda_{h_o}^{i1}(\tau) = \theta(\tau) \ ; \quad \lambda_{\gamma_o}^{i1}(\tau) = \lambda_{\gamma_o}^o(0) + \delta(\tau) \\ &L_o^{i1}(\tau) = 1 + \eta(\tau) \end{aligned}\right\} \tag{19}$$

Substituting these into (15) and using (16) to eliminate $\eta(\tau)$ yields

$$\left.\begin{aligned}
\frac{d\alpha}{d\tau} &= v\beta \\
\frac{d\beta}{d\tau} &= \frac{1}{v}\left(\frac{2}{v^2} - \frac{B_h}{B}\right)\alpha + \frac{1}{v\lambda_\gamma}\,\delta \\
\frac{d\theta}{d\tau} &= \left[\left(\frac{B_h}{B} - \frac{2}{v^2}\right)\left(\lambda_E p_{hL} + \frac{\lambda_\gamma}{v^3}\right) - \lambda_E p_{hh}\right]\alpha - \left(\frac{\lambda_E p_{hL}}{\lambda_\gamma} + \frac{1}{v^3}\right)\delta \\
\frac{d\delta}{d\tau} &= -\frac{\lambda_\gamma}{v}\,\beta - v\theta
\end{aligned}\right\} \tag{20}$$

with

$$\eta = \frac{1}{\lambda_\gamma}\,\delta + \left(\frac{2}{v^2} - \frac{B_h}{B}\right)\alpha \tag{21}$$

Note in these equations that all coefficients are evaluated on the outer solution at $t = 0$, e.g., $v = v_o^o(0)$. From equations (6.19) and (6.35) of [4],

$$\left(\frac{\lambda_E p_{hL}}{\lambda_\gamma} + \frac{1}{v^3}\right) = \frac{1}{v}\left(\frac{2}{v^2} - \frac{B_h}{B}\right) \tag{22}$$

For simplicity, let

$$\left.\begin{aligned}
c &= \frac{1}{v}\left(\frac{2}{v^2} - \frac{B_h}{B}\right) \\
e &= -\frac{\lambda_\gamma}{v}\left(\frac{2}{v^2} - \frac{B_h}{B}\right)^2 - \lambda_E p_{hh}
\end{aligned}\right\} \tag{23}$$

Then (20) may be written

$$\left.\begin{aligned}
\frac{d\alpha}{d\tau} &= v\beta \\
\frac{d\beta}{d\tau} &= c\alpha + \frac{1}{v\lambda_\gamma}\,\delta \\
\frac{d\theta}{d\tau} &= e\alpha - c\delta \\
\frac{d\delta}{d\tau} &= -\frac{\lambda_\gamma}{v}\,\beta - v\theta
\end{aligned}\right\} \tag{24}$$

so that

$$\underline{\underline{G}} = \begin{pmatrix} 0 & v & 0 & 0 \\ c & 0 & 0 & \frac{1}{v\lambda_\gamma} \\ e & 0 & 0 & -c \\ 0 & -\frac{\lambda_\gamma}{v} & -v & 0 \end{pmatrix} \tag{25}$$

The characteristic equation of $\underline{\underline{G}}$ is

$$S^4 + K_2 S^2 + K_1 = 0 \tag{26}$$

where

$$K_1 = -\frac{P_{hh}}{2vBw}\ ;\ K_2 = \frac{2B_h}{B} - \frac{3}{v^2} \tag{27}$$

Thus the eigenvalue criterion for this problem may be stated as

$$\left.\begin{array}{l}\text{either } 4K_1 > K_2^2 \\ \text{or } K_1 > 0 \text{ and } K_2 < 0\end{array}\right\} \quad (28)$$

Note that inspection of (25) and (26) verifies that $\underline{\underline{G}}$ is of the form of (12) and that (13) will be satisfied.

Following the suggestion of Aiken and Lapidus [3], (15) and (16) are now approximated by imposing (17) or (18) on the linearized equations (24). We assume that we have case (i) of (14) and let S_i be any eigenvalue of (26). Set

$$a = |\mathrm{Re}(S_i)| \neq 0 \; ; \quad b = |\mathrm{Im}(S_i)| \neq 0 \quad (29)$$

The general solution of (24) which decays exponentially is then

$$\left.\begin{array}{l}\alpha = A^{(1)} e^{S_1\tau} + A^{(2)} e^{S_2\tau} \\ \beta = \mu_2^{(1)} A^{(1)} e^{S_1\tau} + \mu_2^{(2)} A^{(2)} e^{S_2\tau} \\ \theta = \mu_3^{(1)} A^{(1)} e^{S_1\tau} + \mu_3^{(2)} A^{(2)} e^{S_2\tau} \\ \delta = \mu_4^{(1)} A^{(1)} e^{S_1\tau} + \mu_4^{(2)} A^{(2)} e^{S_2\tau}\end{array}\right\} \quad (30)$$

where

$$S_1 = -a + ib \; ; \quad S_2 = -a - ib \quad (31)$$

In (30), $A^{(1)}$ and $A^{(2)}$ are two arbitrary constants to be determined by the initial conditions. The mode shapes are given by

$$\left.\begin{array}{l}\mu_2^{(i)} = \dfrac{S_i}{v} \\ \mu_3^{(i)} = \dfrac{e}{S_i} - \dfrac{c\lambda_\gamma}{S_i}(S_i^2 - vc) \\ \mu_4^{(i)} = \lambda_\gamma(S_i^2 - vc)\end{array}\right\} \quad (32)$$

The initial conditions on (30) are determined by (17), (18), and (19) as

$$\alpha(0) = h_o - h \text{ and } \beta(0) = \gamma_o \text{ (problem A)} \quad (33)$$

$$\alpha(0) = h_o - h \text{ and } \delta(0) = -\lambda_\gamma \text{ (problem B)} \quad (34)$$

First consider problem A. From (30) and (33), the equations for $A^{(1)}$ and $A^{(2)}$ are

$$h_o - h = A_A^{(1)} + A_A^{(2)}$$

$$\gamma_o = \mu_2^{(1)} A_A^{(1)} + \mu_2^{(2)} A_A^{(2)}$$

which gives, by use of (32),

$$\left.\begin{array}{l}A_A^{(1)} = \dfrac{v\gamma_o - S_2(h_o - h)}{S_1 - S_2} \\ A_A^{(2)} = \dfrac{-v\gamma_o + S_1(h_o - h)}{S_1 - S_2}\end{array}\right\} \quad (35)$$

Putting these in (30) and using (31) yields

$$\left.\begin{aligned} \alpha(\tau) &= e^{-a\tau}\left\{(h_o - h)\cos b\tau + [v\gamma_o + a(h_o - h)]\,\frac{\sin b\tau}{b}\right\} \\ \beta(\tau) &= e^{-a\tau}\left\{\gamma_o\cos b\tau - [av\gamma_o + (a^2 + b^2)(h_o - h)]\,\frac{\sin b\tau}{bv}\right\} \end{aligned}\right\} \tag{36}$$

(Recall that $h = h_o^o(0)$ and $v = v_o^o(0)$ in these equations.) The slopes of α and β evaluated at $t = 0$ are of interest:

$$\left.\begin{aligned} \frac{d\alpha}{d\tau}(0) &= v\gamma_o \\ \frac{d\beta}{d\tau}(0) &= -2a\gamma_o - \frac{a^2 + b^2}{v}(h_o - h) \end{aligned}\right\} \tag{37}$$

The first of these is the value predicted by the state equations for small γ. Thus, for small γ_o, the linear approximation should give good agreement at least near $t \doteq 0$ and as $t \to \infty$ (but probably not at intermediate values of γ because of the steep initial dive characteristic of most MTC trajectories). The second of (37) depends on γ and $(h_o\text{-}h)$ in the correct sense.

For problem B, (30), (34), and (32) give

$$\left.\begin{aligned} A_B^{(1)} &= \frac{-1 - (h_o - h)(S_2^2 - vc)}{S_1^2 - S_2^2} \\ A_B^{(1)} &= \frac{1 + (h_o - h)(S_1^2 - vc)}{S_1^2 - S_2^2} \end{aligned}\right\} \tag{38}$$

and hence

$$\left.\begin{aligned} \alpha(\tau) &= \frac{(h_o - h)e^{-a\tau}}{2ab}\left[2ab\cos b\tau + \left(\frac{1}{h_o - h} - \frac{2}{v^2} + \frac{B_h}{B} + a^2 - b^2\right)\sin b\tau\right] \\ \beta(\tau) &= \frac{(h_o - h)e^{-a\tau}}{2v}\left[\frac{1}{a}\left(\frac{1}{h_o - h} - \frac{2}{v^2} + \frac{B_h}{B} - a^2 - b^2\right)\cos b\tau + \frac{1}{b}\left(-\frac{1}{h_o - h} + \frac{2}{v^2} - \frac{B_h}{B}\right.\right. \\ &\qquad \left.\left. -a^2 - b^2\right)\sin b\tau\right] \end{aligned}\right\} \tag{39}$$

The second of these predicts the value of γ_o when it is left free:

$$\gamma_o = \gamma_o^{i1}(0) = \beta(0) = \frac{1 - (h_o - h)\left(\frac{2}{v^2} - \frac{B_h}{B} + a^2 + b^2\right)}{2av} \tag{40}$$

The solutions of the linearized terminal boundary-layer equations are similar to those of the initial layer. To obtain the terminal solutions from (36) and (39), change $\tau = t/\varepsilon$ to $\tau' = (T-t)/\varepsilon$, v to $-v$, h_o to h_f, and evaluate coefficients at the terminal values of the zeroth-order outer solution.

As mentioned previously, use of the linearized zeroth-order boundary-layer solutions has two major advantages over dealing with the nonlinear equations. First, there is no need for the repetitive, coupled, numerical integrations needed to solve a nonlinear two-point boundary value problem. Second, the integrals on the interval $[0,\infty)$ of the zeroth-order boundary-layer variables which are needed to obtain first-order terms are available in closed form without recourse to numerical integrations to arbitrary cutoff points. In fact, these two advantages mean that using the linear solutions would reduce the computational cost of numerical solutions of the MTC problem to an all but insignificant amount. The question remains, however, as to the numerical validity of the linear solution (investigated in next section).

A possible use of the linear solution is to predict the unknown initial conditions for the nonlinear equations. From (19), (30), (32), and (35), the value of the unknown initial condition in problem A is predicted by the linear solution to be

$$\lambda_{\gamma_o}^{il}(0) = \lambda_{\gamma}\left[1 - 2av\gamma_o - (h_o - h)\left(a^2 + b^2 + \frac{2}{v^2} - \frac{B_h}{B}\right)\right] \tag{41}$$

For γ_o free (problem B), $\lambda_{\gamma_o}^{il}(0) = 0$ and, for this case, (41) necessarily gives the same value for γ_o as does (40). The numerical validity of (41) is also investigated in the following section.

NUMERICAL EXAMPLE

Consider the same example used in [4] and [7]. It is desired to find the minimum time to climb between a Mach number of 0.5 and an altitude of 40,000 ft and a Mach number of 2.0 and an altitude of 80,000 ft for a certain hypothetical, high-performance aircraft. It was determined in [4] that, for the initial boundary layer, $a = 0.0625$ and $b = 0.0738$ so that we do, in fact, have case (i) of (15). Relation (13) gives the eigenvalues as $0.0625 + i.0738$, $-0.0625 - i.0738$, $0.0625 - i.0738$, and $-0.0625 + i.0738$.

The linear solution for a representative problem A ($\gamma_o = -2^{\circ}$) is shown in Figs. 1 and 2. The nonlinear solution is also shown for comparison. The two solutions show fairly good qualitative agreement in both h_o^{il} and γ_o^{il} in that time constants, periods, and initial slopes are about the same for the linear and nonlinear solutions. In particular, the steep dive to $\gamma \approx 60^{\circ}$ is predicted by the linear, as well as by the nonlinear, solution. Quantitatively, however, the agreement is poor. For some values of τ, there is considerable difference in the linear and nonlinear values of h_o^{il} and γ_o^{il}. Further, it is apparent that the integrals of functions of h_o^{il} needed to obtain the first-order solutions will differ considerably. Thus it seems that the linear solutions are of doubtful value in a numerical solution.

The zeroth-order additive composite solutions obtained by use of the linear and nonlinear boundary-layer solutions are compared in Figs. 3 and 4. (The curves labeled linear use linear solutions for both the initial and terminal boundary layers.) These figures again show that the qualitative agreement is good but the quantitative agreement is poor. Note that the nonlinear solution predicts the final value of γ to be 34° while the linear solution predicts 57°. Figure 5 shows the paths in the (M,h) plane. The linear solution actually gives better agreement with the first-order nonlinear solution (considered to be a very good representation of the exact solution) than does the nonlinear. However, although a first-order solution based on the linear zeroth-order boundary-layer solutions has not been worked out, it is anticipated that this would give a poor result.

Figure 6 compares the values of the unknown initial condition, $\lambda^{il}_{\gamma_o}(0)$, as a function of specified γ_o. The linear curve is obtained from (41). Obtaining the nonlinear curve is quite laborious since, for each value of γ_o, the nonlinear ZOBLE's must be repetitively solved for different values of $\lambda^{il}_{\gamma_o}(0)$ until satisfactory boundary-layer solutions are obtained. Considering the ease with which the linear prediction is obtained and the lack of any other good estimating procedures, the agreement evidenced in Fig. 6 is sufficient to make the linear prediction of $\lambda^{il}_{\gamma_o}(0)$ useful as a starting value in the integration of the nonlinear ZOBLE's. Figure 6 indicates that the linear solution predicts the value of γ_o for γ_o free with extreme accuracy. This result, however, is fortuitous and does not hold for other airplane characteristics and other boundary conditions (as evidenced by Fig. 4).

REFERENCES

1. C. R. Hadlock, "Singular Perturbations of a Class of Two Point Boundary Value Problems Arising in Optimal Control," Coordinated Science Lab. Rept. R-481, Univ. Illinois, July 1970.

2. M. I. Freedman and J. L. Kaplan, "Singular Perturbations of Two-Point Boundary Value Problems Arising in Optimal Control," SIAM J. Control Optim., Vol. 14, No. 2, Feb. 1976.

3. R. C. Aiken and L. Lapidus, "An Effective Numerical Integration Method for Typical Stiff Systems," AIChE J., Vol. 20, No. 2, March 1974.

4. M. D. Ardema, "Singular Perturbations in Flight Mechanics," NASA TM X-62,380, Aug. 1974, revised Oct. 1975.

5. R. E. O'Malley, Jr. and C. F. Fung, "The Singularly Perturbed Linear State Regulator Problem, II," SIAM J. Control, Vol. 13, No. 2, Feb. 1975.

6. H. J. Kelley, "Singular Perturbations for a Mayer Variational Problem," AIAA J., Vol. 8, No. 6, June 1970.

7. M. D. Ardema, "Solution of the Minimum Time-to-Climb Problem by Matched Asymptotic Expansions," AIAA J., Vol. 14, No. 7, July 1976.

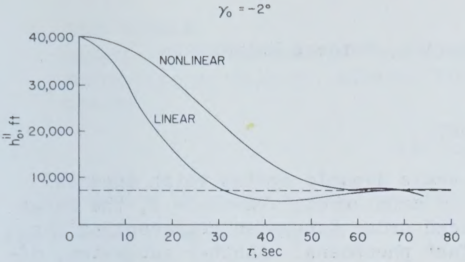


Fig. 1.- Boundary-layer solutions for h.

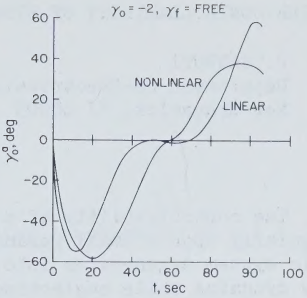


Fig. 4.- Composite solutions for γ.

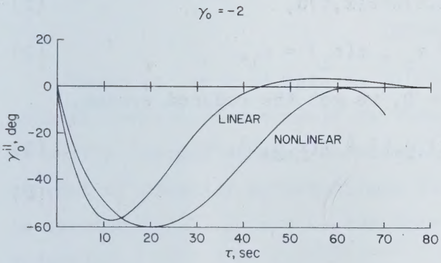


Fig. 2.- Boundary-layer solutions for γ.

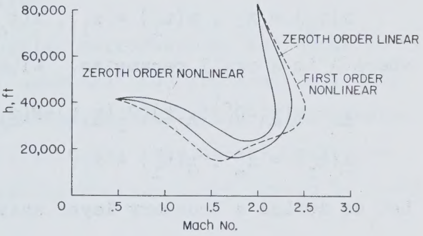


Fig. 5.- Comparison of paths.

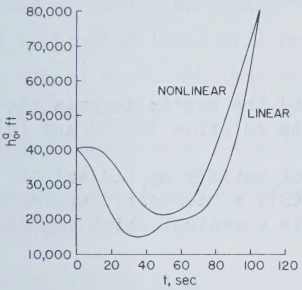


Fig. 3.- Composite solutions for h.

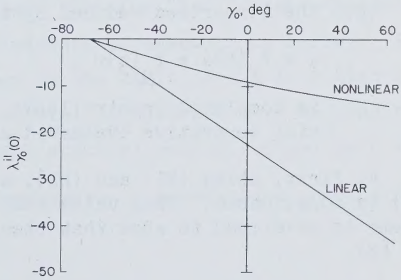


Fig. 6.- Prediction of $\lambda_\gamma(0)$.

ON THE CONTROLLABILITY OF SINGULARLY PERTURBED NONLINEAR SYSTEMS

P. SANNUTI
Department of Electrical Engineering, Rutgers University
New Brunswick, NJ 08903

ABSTRACT

The controllability of a large scale dynamic system which depends singularly upon a small parameter λ is considered. When $\lambda = 0$, the large scale system degenerates into a reduced order subsystem representing its slow dynamics while neglecting the fast phenomena. Another subsystem, often called a boundary layer system, represents the fast dynamics. In this paper sufficient conditions are established under which the controllability of the overall large scale system is inferred from the same property of the two subsystems.

MAIN RESULT

Consider a large scale dynamic system,

$$\dot{x} = g(x,z,u,t) \ , \ \lambda\dot{z} = G_1(x,t)+D(x,t)z+E(x,t)u, \tag{1}$$

$$x(t_o) = x_o \ , \ x(t_1) = x_1 \ , \ z(t_o) = z_o \ , \ z(t_1) = z_1, \tag{2}$$

where λ is a small parameter. With $\lambda = 0$, we get the reduced system,

$$\dot{\underline{x}} = g(\underline{x},-D^{-1}(\underline{x},t)(G_1(\underline{x},t)+E(\underline{x},t)\underline{u}),\underline{u},t) \triangleq f(\underline{x},\underline{u},t), \tag{3}$$

$$\underline{x}(t_o) = x_o \ , \ \underline{x}(t_1) = x_1. \tag{4}$$

Let us define a boundary layer system,

$$\frac{dz}{d\rho} = D(x,t)z + E(x,t)u. \tag{5}$$

Here x and t are considered as parameters and ρ is an independent time variable. Under the following hypotheses (H1)-(H3), it is shown that, for λ sufficiently small, a control $u(t)$ exists such that the solution of (1) satisfies (2).

(H1) A control $\underline{u}(t)$ exists satisfying the reduced system (3) and (4).
(H2) The linear system (5) is controllable for all x and t.
(H3) The linearized reduced system,

$$\alpha = f_{\underline{x}}(t)\alpha + f_{\underline{u}}(t)\nu \ , \tag{6}$$

is completely controllable. In (6) the suffix denotes the partial derivative evaluated along the solution of (3) and (4).

At first, using (H1) and (H2), a control satisfying (1) and (2) within $O(\lambda)$ is constructed. Then using (H2) and (H3), a successive approximation scheme is developed to show that there exists a control which satisfies (1) and (2).

SINGULARLY PERTURBED OPTIMAL CONTROL PROBLEMS

PAUL BINDING
Department of Mathematics and Statistics
University of Calgary, Alberta T2N 1N4
Canada

ABSTRACT

A review is given of some factors affecting convergence as $\varepsilon \downarrow 0$ for the problem of minimising $c\left(T_\varepsilon\right) - c\left(0\right)$ given

$$\dot{c} = e(x,y,u,t), \quad \dot{x} = f(x,y,u,t), \quad \varepsilon\dot{y} = g(x,y,u,t) \quad \text{on } [0,T_\varepsilon]$$

where variable end conditions are imposed on x for $\varepsilon \geq 0$ and on y for $\varepsilon > 0$, and where controls u take values in a bounded set $\mathsf{U} \subseteq \mathbb{R}^m$. The factors studied include smoothness of end conditions, differential equations and optimal controls; nontangency to end sets; stability of g in y; and nonsingularity of optimal controls.

INTRODUCTION

There are several ways in which singular perturbations can occur in optimal control, for example by viewing "abnormal" [4] problems as limits of nearly "normal" ones. An example is "cheap control" [5],[7]; a general concept of singular perturbations for differential (and other) equations is treated in [3]. Here we consider singular perturbations caused by order reduction.

Specifically, the general problem for $\varepsilon \geq 0$ is to minimise the scalar $c\left(T_\varepsilon\right)$ given $c(0) = 0$, $x(0) \in X_0 \subseteq \mathbb{R}^n$, T_ε being fixed or else the least time such that $x\left(T_\varepsilon\right) \in X_1 \subseteq \mathbb{R}^n$. c, x and y satisfy the differential equations displayed above and for $\varepsilon > 0$ we also impose $y(0) \in Y_0 \subseteq \mathbb{R}^k$. The case $\varepsilon = 0$ is called "reduced" and involves a "slow" state variable x satisfying an nth order differential equation, provided that the "fast" state variable y can be eliminated (this will be a consequence of assumptions below). In a formal sense, then, the reduced problem is the limit of the $(n + k)$th order "full" problems with $\varepsilon > 0$ and it is natural to ask what sort of conditions guarantee "convergence" of optimal cost, state(s) and/or control as $\varepsilon \downarrow 0$.

We consider a variety of modes and orders of convergence, depending on three pairs of conditions.

A End conditions on x

Typically these are specified by equations or inequalities (and could have ε dependence, although this adds no real novelty, so we shall suppress it). Perturbation of trajectories moves the end points and bounds are needed on this movement. Classically they are furnished by (i) nontangency to, and (ii) smoothness of, the end sets X_i (which are frequently specified by equations and inequalities). Differential dependence of end points on initial points then results, but we shall also consider less smooth dependence.

B Conditions on the differential equations

Again we have equations and inequalities (defining U) and again one type of condition needed is (i) smoothness. Another is (ii) stability relative to the end conditions on y. For simplicity we have left $\odot(T_\varepsilon)$ free; it suffices then if g is "stable" in y. The sense will be made more precise below, but we expect perturbations in y to be damped out, say with time constant $0(1/\varepsilon)$. In principle, closed loop control could be used for stabilisation, but for simplicity we shall consider only open-loop stability.

C Conditions on the optimal control functions

The first condition is again (i) smoothness. When the reduced optimal control jumps, the situation parallels that for the initial condition on y, with boundary layers at least when $\varepsilon > 0$. The full y_ε is a smooth function starting in Y_0; assuming that g is smoothly invertible in y, the reduced y_0 starts where $0 = g(x,y,u,t)\big|_{t=0}$ and jumps whenever u_0 does. Even if u_0 is continuous, $0(1/\varepsilon)$ dynamics may not be able to keep up, and our best estimates on $y_\varepsilon - y_0$ may have to be "integrated."

We shall also need to know how far the reduced optimal control moves under perturbation. Depending on the convergence results sought, this (ii) "normality" condition may vary from one of existence type to differential nonsingularity (i.e. classical embeddability in an extremal field).

The object now is to present examples illustrating these six types of condition and to discuss the corresponding effects on convergence.

EXAMPLE 1

Let $\dot{c} = 1$, $\ddot{x} = y$, $\varepsilon\dot{y} = u - y$, $|u(t)| \leq 1 = \dot{x}(0)$, $x(0) = -\frac{1}{2}$, $y(0) = x(T_\varepsilon) = \dot{x}(T_\varepsilon) = 0$.

When $\varepsilon = 0$, this is a standard textbook problem, with $u_0 \equiv -1$ and $T_0 = 1$. Singular perturbation drives $x_\varepsilon(T_0)$ to $0(\varepsilon)$, and

$$T_\varepsilon = c(T_\varepsilon) = T_0 + 0(\varepsilon^{\frac{1}{2}}).$$

Thus a problem that could scarcely be "smoother" does not admit a power series expansion in integer powers of ε. This failure of Ai also occurs with end condition $x(T_\varepsilon) = 0 \geq \dot{x}(T_\varepsilon)$ but not, of course, with $x(T_\varepsilon) = 0$ alone. A similar example (of Aii) is provided by $x(T_\varepsilon) + |\dot{x}(T_\varepsilon)| \leq 0$.

Note that these examples are really nothing to do with *singular* perturbations, or even with *control* per se. They depend on end point variation, and this frequently occurs with singular perturbations. More drastic versions, where the reduced solution is perturbed into a full problem which has *no* solutions satisfying the end conditions, are easily given [1].

On the positive side, continuous variation of x end points (on d.e. right sides) will give continuous dependence of cost, state, etc. on ε, under enough conditions of type B and C. Lipschitz variation gives $O(\varepsilon)$ estimates between full and reduced costs, states, etc. C^1 variation allows an explicit "ε term" to split off, starting a series expansion, and so on.

EXAMPLE 2

Let $\dot{c} = y$, $\varepsilon\dot{y} = u - y + \varphi$, $y(0) = 0 \leq u(t) \leq 1 = T$, $\varphi\colon [0,1] \to \mathbb{R}$.

Obviously $u \equiv 0$ is optimal, and could be absorbed into the definition of φ, illustrating that this is a d.e. phenomenon, not really a *control* one. Bi is the relevant condition. If $\varphi(t) = \operatorname{sgn}(t - \tfrac{1}{2})$ then we have a boundary layer at $t = \tfrac{1}{2}$. If $\varphi \in L_1$, the boundary layer may be "perpetual" but it can be shown [1] that $y_\varepsilon \to y_0$ in L_1 as $\varepsilon \downarrow 0$. Likewise the optimal cost $c_\varepsilon(1) \to c_0(1)$.

We can do better if φ is smoother. For example $\varphi \in AC$ gives $y_\varepsilon = y_0 + O(\varepsilon)$ in L_1, but there may still not be a useful *uniform* estimate outside the initial boundary layer. For instance $\varphi(t) = 1/\log|\tfrac{1}{2} - t|$ gives

$$|(y_\varepsilon - y_0)(\tfrac{1}{2} - \varepsilon)| = O(1/|\log \varepsilon|)$$

which is $O(\varepsilon^\alpha)$ for *no* $\alpha > 0$. Lipschitz φ does allow a $O\left(e^{-t/\varepsilon}\right)$ initial layer plus a uniform $O(\varepsilon)$ term. Whether the latter can be decomposed into a multiple of ε ("ε term") plus a smaller term depends on further smoothness. For example $\dot{\varphi} \in AC$ allows a uniform $o(\varepsilon)$ "smaller term".

These statements are easy consequences for Ex. 2 of

$$y_\varepsilon(t) = \int_0^t \varphi(s) e^{(s-t)/\varepsilon} ds/\varepsilon$$

and integrations by parts. They are, however, representative of general nonlinear systems and extend to allow $O(\varepsilon^\alpha)$ estimates if

$$\int_0^{T-s} |\varphi(s+t) - \varphi(t)|\,dt = 0(s^\alpha).$$

This has some connection with the theory of interpolation spaces [3].

Asymptotic expansions are often developed for d.e. and control problems. They depend, essentially, on Taylor expansions of the d.e. right sides. Thus failure of some order of differentiability (in x,y,u or t) may cause failure of asymptotic expansions to that order.

EXAMPLE 3

Let $\dot{c} = y$, $\varepsilon\dot{y} = \alpha y^\beta + u$, $0 \le u(t) \le 1 = y(0) = T$.

Again this generates d.e. (not control per se) phenomena, Bii being the relevant condition. Obviously $u \equiv 0$ is optimal, and if $\alpha > 0$ then the d.e. for y "blows up" as $\varepsilon \downarrow 0$. $\alpha < 0$, $\beta = 1$ is essentially Ex. 2 and $y_\varepsilon(t) = e^{-\alpha t/\varepsilon}$. We can say that y is "strongly" stable, and evidently $y_\varepsilon = 0(\varepsilon)$ in L_1.

When $\alpha < 0 < \beta < 1$, stability is no problem, but nondifferentiability may cause difficulties with expansions, at least in coupled systems. When $\beta > 2$, y is stable, but not "strongly" so. In fact,

$$y_\varepsilon(t) = (1 + (1-\beta)\alpha t/\varepsilon)^{1/(1-\beta)} = \varepsilon^{1/(\beta-1)}(\varepsilon + (1-\beta)\alpha t)^{1/(1-\beta)}$$

so $y_\varepsilon(\varepsilon) = 0(1)$, illustrating the usual nonuniform convergence near $t = 0$. Even in L_1, we can only manage a less than unit power of ε, specifically $\|y\|_1 = 0\left(\varepsilon^{1/(\beta-1)}\right)$.

This example is again representative of more general situations - the state stability effect depends on the behaviour of g in y. It is also sometimes possible to treat borderline nonstable cases like

$$\varepsilon^2\ddot{y} + y = 0, \quad \dot{y}(0) = 0, \quad y(0) = T = 1.$$

In this case $y(t) = \cos(t/\varepsilon)$ which is weakly (but not strongly) convergent to zero in L_1. Thus $\dot{c} = y$ gives $|c_\varepsilon(1) - c_0(1)| = 0(\varepsilon)$. On the other hand, a non-weakly-continuous functional, given by say $\dot{c} = y^2$, allows only $|c_\varepsilon(1) - c_0(1)| = 0(1)$.

EXAMPLE 4

Let $\dot{c} = uy$, $\varepsilon^2\ddot{y} + 2\varepsilon\dot{y} + 2y = 0$, $|u(t)| \le 1 = \dot{y}(0) = T$, $y(0) = 0$.

In this case $y_0 = 0$ so all controls are optimal, and $c_0(1) = 0$. The full problem, however, possesses solutions

$$y_\varepsilon(t) = e^{-t/\varepsilon} \sin(t/\varepsilon), \quad u_\varepsilon(t) = -\operatorname{sgn} y_\varepsilon(t) = -\operatorname{sgn} \sin(t/\varepsilon).$$

This produces a phenomenon similar to that of Ex. 2 with $\varphi \in L_1$ but for a reason that really has something to do with optimal control this time.

Evidently the $0(1/\varepsilon)$ dynamics cannot keep up with control switches at $0(1/\varepsilon)$ intervals - a Ci difficulty. The set of optimal controls u_ε is not even L_1 compact. (Actually, convergence of $c_\varepsilon(T)$ to $c_0(T)$ can be established because the y_ε are L_1 compact [1]). The comments below Ex. 2 on smoothness apply equally here to (open loop) control of course.

It can be shown, under appropriate conditions of type A and B, that convergence of c, x (in AC, hence uniformly) and y (in L_1) holds if the u_ε are compact in L_1 [1]. In particular, this is so if the u_ε have uniformly bounded variation. Assume, further, that the "singular" parts of the u_ε are purely discontinuous, and the AC parts are Lipschitz. Then under appropriate conditions of type A and B, "ε terms" can be split off for a given control, and even for optimal control if the discontinuities vary smoothly in ε and the Lipschitz constant is independent of ε [2].

EXAMPLE 5

Let $\dot{c} = y^2$, $\varepsilon\dot{y} = u - y$, $u(t) = \pm 1$, $y(0) = 0$, $T = 1$.

The reduced problem is $\dot{c}_0 = u^2$, and obviously any control is optimal, with $c_0(T) = 1$. The full problem has no minimum although the infimal cost is zero, obtained with the weak ("relaxed") control $u_\varepsilon \equiv 0$.

The difficulty is of type Cii - the reduced optimal control is not embeddable at all (let alone smoothly) into a field of full optimal controls. Notice the lack of convexity in the set of allowable derivatives, a typical reason for nonexistence of optimal controls.

One can look at this example from a viewpoint similar to that for Ex. 4. The relaxed control $u_\varepsilon \equiv 0$ can be approximated say by $v_\varepsilon(t) = \operatorname{sgn} \sin(t/\varepsilon)$ - note that $v_\varepsilon \to 0$ weakly in L_1 - and it is easy to check that the corresponding costs do converge to zero with ε.

EXAMPLE 6

Let $\dot{c} = x^2$, $\dot{x} = y$, $\varepsilon\dot{y} = u - y$, $|u(t)| \leq 1 = x(0)$, $y(0) = 0$, $T = 2$.

The reduced d.e. are $\dot{c}_0 = x_0^2$, $\dot{x}_0 = u$ with optimal solution

$$u_0 \equiv -1, \quad x_0(t) = 1 - t \qquad \text{on } [0,1]$$
$$u_0 \equiv 0 \equiv x_0 \qquad \text{on } [1,2],$$

giving cost $c_0(T) = 1/3$. Observe that u_0 is nonsingular on $[0,1[$ and singular on $[1,2]$.

When $\varepsilon > 0$, the optimal control shatters into a "nonanalytic junction" [6] joining nonsingular and (2nd order) singular arcs. Observe that, with infinitely many jumps of magnitude 2, the full optimal controls are not even of bounded variation, so the comments below Ex. 4 do not apply. Although $O(1/\varepsilon)$ dynamics obviously cannot keep up, all is not lost: the reduced optimal control u_0 is only $O(\varepsilon^3)$ suboptimal for the full model. This is because the unruly full optimal control behaviour occurs precisely because of the cost insensitivity that singularity implies.

Actually Exs 4, 5 and 6 all display singularity. Asymptotic expansions can be guaranteed by differential nonsingularity - strong Legendre and Dresden (corner) conditions giving embeddability of the reduced optimal trajectory into a field of similar extremals [2]. Singular reduced solutions (and tangency at end conditions) can sometimes be treated in terms of fractional powers of ε - cf. Ex. 1.

REFERENCES

[1] P. Binding, "Singularly perturbed optimal control problems": I Convergence, SIAM J. Control and Optimization 12 (1976), 591-612.

[2] P. Binding, Ibid: II Estimates (to appear).

[3] P. Binding, "Singular perturbations via monotonicity methods" (to appear)

[4] G. Bliss, "Lectures on the Calculus of Variations," U. of Chicago Press, 1946.

[5] J. Lions, "Perturbations Singulières dans les Problèmes aux Limites et en Contrôle Optimal", Lecture Notes in Mathematics 323, Springer, 1970.

[6] J. McDanell, W. Powers, "Necessary conditions for joining optimal singular and nonsingular subarcs," SIAM J. Control 9 (1971), 161-173.

[7] R. O'Malley, "A more direct solution of the nearly singular linear regulator problem," SIAM J. Control and Optimization (to appear).

THE TIME-OPTIMAL CONTROL OF A CLASS OF NONLINEAR SINGULARLY PERTURBED SYSTEMS*

S. H. JAVID
Coordinated Science Laboratory and Department of Electrical Engineering
University of Illinois, Urbana, Illinois 61801

ABSTRACT

The time-optimal control of systems made of nonlinear plants with fast linear actuator dynamics is shown to consist of two parts. The first part is primarily dependent on the dynamics of the nonlinear plant and second part is primarily dependent on the fast actuator dynamics. Based on this characterization of the control a near optimal control is proposed.

INTRODUCTION

Recently a number of results [1-5] have been obtained for the time-optimal control of linear time invariant singularly perturbed systems. Using the results of [2] an iterative method has been developed [4,5] which iterates for the time-optimal control from a near optimal control. This paper extends the work of [1-3] to treat a "quasi-linear" system

$$\dot{x} = f(x,t) + A_{12}(t)z + B_1(t)u \tag{1}$$

$$\mu\dot{z} = A_2(t)z + B_2(t)u \tag{2}$$

where x, z and u are n, m and r vectors respectively, μ is a small positive scalar, the components of u are constrained

$$|u_i| \leq 1, \qquad i = 1,\ldots,r$$

and the eigenvalues of $A_2(t)$ satisfy $\mathrm{Re}(\lambda_i) < -\sigma < 0$ for all $t > t_0$.

The control, which drives the state of (1),(2), (x',z') from the initial state (x_0',z_0') at time $t_0 = 0$ to the final state (x_f',z_f') in minimum time T^*, is denoted $u^*(t)$. The resulting optimal trajectory is $(x^{*\prime}, z^{*\prime})$. Due to the presence of μ, x(t) and z(t) are referred to as the slow and fast states of (1),(2). The fast dynamics of (1),(2) may be neglected by setting μ to zero. The result is the reduced order system

$$\dot{\bar{x}} = f(\bar{x},t) + B_0(t)\bar{u} \tag{3}$$

$$\bar{z} = -A_2(t)^{-1} B_2(t)\bar{u} \tag{4}$$

---

*This work was supported in part by the National Science Foundation under Grant ENG 74-20091, in part by the Energy Research and Development Administration under Contract U.S. ERDA E(49-18)-2088 and in part by the U.S. Air Force under Grant AFOSR 73-2570.

where $B_0(t) = B_1(t) - A_{12}(t)\, A_2(t)^{-1} B_2(t)$. We may define the reduced time-optimal problem as that of finding the control $\bar{u}^*(t)$ which steers $\bar{x}$ of (3) from x_0 to x_f in minimum time $\bar{T}^*$. The resulting optimal trajectory is $\bar{x}^*$.

For the reduced-order system to provide an approximation for the x dynamics in (1),(2) subsystem (2) must be asymptotically stable. We know [6] that this is true for μ small enough, since the real parts of the eigenvalues of $A_2(t)$ are smaller than a negative constant $\forall t > t_0$. Suppose that $\bar{u}^*(t)$ is applied to (1),(2). The value $x(\bar{T}^*)$ will be shown to be within $O(\mu)$ of x_f and $z(\bar{T}^*)$ will be different from z_f. Denote the control which steers z of (2) from $z(\bar{T}^*)$ to z_f in minimum time t_f^*, $u_f(t)$. This paper shows that the control which is made of $\bar{u}^*(t)$ followed by $u_f(t)$ is a near time-optimal control for (1),(2) in the sense that it steers (x',z') in a near time-optimal time $T^* + O(\mu)$.

In the next section some preliminary results are presented following which the near time-optimal control is developed. Some implications of this control are discussed and an example is presented.

SLOW AND FAST CONTROL

We first present some preliminary results. As mentioned above subsystem (2) must be stable. We know [6] that for any $\varepsilon_0 > 0$ there exists a μ_0 such that for $\mu \in (0,\mu_0]$ the state transition matrix of (2), $\phi_2(t,t_0)$ satisfies

$$\frac{1}{\mu}\, \phi_2(t,t_0) = O(\mu) \tag{5}$$

for $t > t_0 + \varepsilon_0$. Similarly the state transition matrix of (14), $\phi_2'(t,T^*)$, satisfies

$$\frac{1}{\mu}\, \phi_2'(t,T^*) = O(\mu) \tag{6}$$

for $t < T^* - \varepsilon_0$. Furthermore ε_0 decreases as μ_0 decreases. For the remainder of the paper we assume that $\mu \in (0,\mu_0]$.

Now suppose $\bar{u}^*(t)$ is applied to (1),(2). We are interested in the resulting value of x(t). Let x(t) satisfy

$$x(t) = \bar{x}^* + \omega(t), \qquad \omega(t_0) = 0 \tag{7}$$

then

$$\dot{\omega} = f_x(\hat{x}(t),t)\, \omega + A_{12}(t)(z-\bar{z})$$

where $\hat{x}(t)$ is $\forall t$ the point satisfying the Taylor Theorem for

$$f(x,t) = f(\bar{x}^*,t) + f_x(\hat{x},t)\, \omega(t)$$

Since $\hat{x}$ is well defined we may write

$$\dot{\omega} = A_1(t)\,\omega + A_{12}(t)(z-\bar{z})$$

which has the solution

$$\omega(t) = \int_{t_0}^{t} \phi_1(t,\tau)\, A_{12}(\tau)(z(\tau) - \bar{z}(\tau))d\tau$$

Evaluating $z(t) - \bar{z}(t)$ on intervals $[t_i, t_{i+1})$ where $\bar{u}^*(t) = u_c$ and assuming that ϕ_1 and A_{12} are bounded, $\omega(t)$ satisfies

$$\begin{aligned} \omega(t) = \mu[&A_{12}(t)A_2^{-1}(t)\phi_2(t,t_i) + \phi_1(t,t_i)A_{12}(t_i)A_2^{-1}(t_i)] \cdot \\ &(z(t_i) + A_2^{-1}(t_i)B_2(t_i)u_c) + 0(\mu) \end{aligned} \tag{8}$$

which implies $\omega(t) = 0(\mu)$. Thus from (7) and (8)

$$x(\bar{T}^*) = x_f + 0(\mu)\ . \tag{9}$$

Next suppose we consider $\bar{u}^*(t)$ and a control $\hat{u}(t)$ each of whose components differ from those of $\bar{u}^*(t)$ on a finite number of time intervals of length $0(\mu)$. Since $x(t)$ equals $\bar{x}^*(t) + 0(\mu)$ when $\bar{u}^*(t)$ is applied to (1), (2) and $\hat{u}(t)$ will further perturb $x(t)$ by $0(\mu)$, applying $\hat{u}(t)$ to (1), (2) and $\bar{u}^*(t)$ to (3) will result in

$$\hat{x}(t) - \bar{x}^*(t) = 0(\mu) \tag{10}$$

where $\hat{x}(t)$ is the result of applying $\hat{u}(t)$ to (1), (2). As a result of (10) the state transition matrices of the two differential equations

$$\dot{\bar{p}} = -\ f_x(\bar{x}^*,t)'\bar{p}$$

and

$$\dot{\hat{p}} = -\ f_x(\hat{x},t)'\hat{p}$$

will satisfy

$$\bar{\phi}(t,T) = \hat{\phi}(t,T) + 0(\mu) \tag{11}$$

when $f_x(x,t)$ is bounded.

For (1), (2) the Hamiltonian defined

$$\begin{aligned} H = 1 + p'f(x,t) + p'A_{12}(t)z + p'B_1(t)u + \frac{1}{\mu}\, q'A_2(t)z \\ + \frac{1}{\mu}\, q'B_2(t)u \end{aligned}$$

and the optimal control is

$$u^*(t) = -\mathrm{SGN}[B_1(t)'p + \frac{1}{\mu}\, B_2(t)'q] \tag{12}$$

where the costates p and q satisfy

$$\dot{p} = -f_x(x^*,t)'p \tag{13}$$

$$\dot{q} = -A_{12}(t)'p - \frac{1}{\mu}\, A_2(t)'q\ . \tag{14}$$

Since x* is a well defined trajectory we may write (13) as

$$\dot{p} = -A(t)'p \tag{15}$$

which has the solution

$$p(t) = \phi(t,T^*)p_f. \tag{16}$$

As a result the solution to (14) may be expressed

$$q(t) = \phi_2'(t,T^*)q_f - \int_{T^*}^{t} \phi_2'(t,\tau)A_{12}(\tau)'\phi(\tau,T^*)p_f d\tau. \tag{17}$$

where $\phi_2'(t,T^*)$ is the state transition matrix of $\mu\dot{z} = -A_2(t)'z$. Substituting (16) and (17) into (12) and integrating by parts we find the expression for the time-optimal control of (1),(2).

$$\begin{aligned} u^*(t) = -\mathrm{SGN}[&B_0(t)'\phi(t,T^*)p_f + \frac{1}{\mu} B_2(t)'\phi_2'(t,T^*)q_f \\ &+ B_2(t)'\phi_2'(t,T^*)A_2(T^*)'^{-1}A_{12}(T^*)'p_f + 0(\mu)]. \end{aligned} \tag{18}$$

Next we find the reduced time-optimal control $\overline{u}^*(t)$. For (3) the Hamiltonian

$$\overline{H} = 1 + \overline{p}'f(\overline{x},t) + \overline{p}'B_0(t)\overline{u} \tag{19}$$

and the time-optimal control is

$$\overline{u}^*(t) = -\mathrm{SGN}[B_0(t)'\overline{p}] \tag{20}$$

where

$$\dot{\overline{p}} = -f_x(\overline{x}^*,t)'\overline{p} = -\overline{A}(t)'\overline{p}. \tag{21}$$

The solution of (21) is

$$\overline{p}(t) = \overline{\phi}(t,\overline{T}^*)\overline{p}_f \tag{22}$$

and thus

$$\overline{u}^* = -\mathrm{SGN}[B_0(t)'\overline{\phi}(t,\overline{T}^*)\overline{p}_f]. \tag{23}$$

The SGN function for u*(t), as in (18), operates on two types of terms, one dependent on $\phi_2'(t,T^*)$ and the other similar, in appearance at least to that in (23). From (6) we know that for $t \in [t_0,T^*-\varepsilon_0)$, where ε_0 can be made arbitrarily small by decreasing μ, the terms of the first type are $0(\mu)$. Thus on this interval we may rewrite (18) as

$$u^*(t) = -\mathrm{SGN}[B_0(t)'\phi(t,T^*-\varepsilon_0)\hat{p}_f + 0(\mu)] \tag{24}$$

where $\hat{p}_f = \phi(T^*-\varepsilon_0,T^*)p_f$. The two terms which depend on $\phi_2'(t,T^*)$ have their primary effect on the interval $[T^*-\varepsilon_0,T^*)$.

Suppose that we are not interested in controlling z. Then $q_f = 0$ but the other ϕ_2' dependent term, $B_2(t)'\phi_2'(t,T^*)A_2(T^*)'^{-1}A_{12}(T^*)'p_f$, may create some fast switchings. Clearly as ε_0 decreases these fast switchings will move x very little in (1),(2) and therefore $x^*(T^*-\varepsilon_0)$ must be $0(\mu)$ close to x_f. Note that this is precisely what $\bar{u}^*(t)$ accomplishes as demonstrated by (9), which suggests that on $[t_0,T^*-\varepsilon_0)$ we may approximate $u^*(t)$ by $\bar{u}^*(t)$, the time-optimal control for the reduced system. The fast dynamics for μ small enough change the phase trajectories of the reduced system by bending them at switchings as z rapidly approaches its equilibrium $\bar{z}$. Thus the control sequence of $u^*(t)$ for $t < T^*-\varepsilon_0$ should be the same as that of $\bar{u}^*(t)$ but with the switchings $0(\mu)$ changed, provided the zeros of each component of $B_0(t)'\bar{\phi}(t,\bar{T}^*)\bar{p}_f$ in (23) are simple. These are precisely the conditions for $\bar{\phi}(t,\bar{T}^*) = \phi(t,T^*-\varepsilon_0) + 0(\mu)$ as in (11), which further suggests that as $\mu \to 0$, $\hat{p}_f \to \bar{p}_f$.

When $u^*(t)$ is to control both x and z, q_f is not zero. However, it creates new switchings only on $[T^*-\varepsilon_0,T^*)$. Since μ_0 has been chosen to make (2) asymptotically stable we may interpret $u^*(t)$ as follows. For $t < T^*-\varepsilon_0$ the optimal control steers x to $0(\mu)$ near to x_f, then it rapidly steers z to z_f. At time $\bar{T}^*$ the control $\bar{u}^*(t)$ has steered z to $z(\bar{T}^*)$ which is not z_f. The control $u_f(t)$ which time optimally steers z from $z(\bar{T}^*)$ to z_f in time t_f^* will approximate $u^*(t)$ on $[T^*-\varepsilon_0,T^*)$ in the sense that it will leave x $0(\mu)$ near to x_f. The following lemma summarizes this analysis.

Lemma: The control $\bar{u}^*(t)$, the time-optimal control for the reduced order system (3), steers x to an $0(\mu)$ neighborhood of x_f and z to $z(\bar{T}^*)$. The control $u_f(t)$ time-optimally steers z of (2) from $z(\bar{T}^*)$ to z_f in time t_f^*. There exists a μ^* such that for $\mu \in (0,\mu^*]$ the control consisting of $\bar{u}^*(t)$ followed by $u_f(t)$ is near time-optimal in the sense that $x(\bar{T}^* + t_f^*)$ is $0(\mu)$ close to x_f, $z(\bar{T}^* + t_f) = z_f$ and $\bar{T}^* + t_f^* = T^* + 0(\mu)$.

DISCUSSION

As $\mu \to 0$, the time t_f^* becomes shorter and $x(\bar{T}^*)$ comes closer to x_f while $\bar{T}^*$ remains constant. Furthermore as t_f^* becomes smaller $x(\bar{T}^* + t_f^*)$ remains closer to x_f. Thus for systems in the form (1),(2) the near time-optimal control improves for smaller values of μ.

We might ask: how can we adjust this near time-optimal control so that we get a better approximation of $u^*(t)$? One approach often used with singularly perturbed systems is that of asymptotic expansions to reduce the error in x $O(\mu^2)$ or less as in [2,3]. Another approach which may prove itself useful to this problem is the use of iterative methods as in [4,5].

We note that on the time interval $[T^*-\varepsilon_0, T^*]$ the optimal control contains two parts dependent on $\phi_2'(t,T^*)$. One part involves p_f and the other q_f. Either approach used would have to take this fact into account in seeking the time-optimal control. The great beauty of this near time-optimal control is that it can be found entirely through calculations on lower order subsystems (3) and (2). Suppose we have a system such as (1), (2) for which we would like a near time-optimal control. We calculate the control and see whether the performance of $\bar{u}^*(t)$ followed by $u_f(t)$ is satisfactory. If it is we can say that μ was small enough to satisfy the lemma.

EXAMPLE

Consider the system

$$\begin{aligned} \dot{x} &= -\sin x + \frac{1}{2} z_1 + \frac{1}{2} z_2 \\ \mu\dot{z}_1 &= -z_1 + u \\ \mu\dot{z}_2 &= -2z_2 + 2u \end{aligned} \qquad (25)$$

where $\mu = 0.1$. Let the initial and final state for (25) be $(\pi/2, 0.7, -0.8)$ and $(0,0,0)$ respectively. By setting $\mu = 0$ in (25) we find the reduced order system

$$\dot{\bar{x}} = -\sin \bar{x} + \bar{u} \ . \qquad (26)$$

The reduced time-optimal control

$$\bar{u}^*(t) = -1 \qquad t \in [0,1)$$

steers $\bar{x}$ of (26) from $\pi/2$ to 0. When this control is applied to (25) (x, z_1, z_2) is steered to $(.0679, -1.00, -1.00)$. This is acceptable if we are not interested in controlling z. For instance if the z variables represent actuator dynamics the final position of the actuator may not be important. But suppose we are interested in bringing the actuator to zero: Then we find the control

$$u_f(t) = \begin{cases} +1 & t \in [1, 1.1098) \\ -1 & t \in [1.1098, 1.139) \end{cases}$$

which steers (z_1, z_2) from $(-1, -1)$ to $(0,0)$. When $\bar{u}^*(t)$ followed by

$u_f(t)$ is applied to (25) the state is steered to (.0658, 0.00, 0.00). Thus x is within 5% of its final state, z is at its final state and the control has been found while only considering reduced order systems.

REFERENCES

1. Collins W. D., "Singular Perturbations of Linear Time-Optimal Control," pp. 123-136 in Recent Math. Developments in Control, D. J. Bell, Ed., Academic Press, N.Y., 1973.
2. Kokotovic P. V., Haddad A. H., "Controllability and Time-Optimal Control of Systems with Slow and Fast Modes," IEEE Trans. on Automatic Control, Vol AC-20, No. 1, 1975.
3. Kokotovic P. V., Haddad A. H., "Singular Perturbation of a Class of Time-Optimal Controls," IEEE Trans. on Automatic Control, Vol. AC-20, No. 1, 1975.
4. Javid S. H., "A Recursive Algorithm for Time-Optimal Control of Singularly Perturbed Systems," Proceedings of Thirteenth Annual Allerton Conference on Circuit and System Theory, (839-847), October 1975.
5. Javid S. J., Kokotovic P. V., "A Decomposition of Time Scales for Iterative Computation of Time-Optimal Controls," to be published in Journal of Optimization Theory and Applications.
6. Haddad A. H., "Linear Filtering of Singularly Perturbed Systems," IEEE Trans. on Automatic Control, Vol. AC-21, No. 4, 1976.

STABILITY AND CONTROL OF STOCHASTIC SYSTEMS WITH WIDE-BAND NOISE DISTURBANCES

G.L BLANKENSHIP
Systems Engineering Department
Case Western Reserve University
Cleveland, Ohio 44106

G.C. PAPANICOLAOU
Courant Institute
New York University
New York, New York 10012

ABSTRACT

For dynamical systems with external influences which are approximately white noise (wide-band noise) we show that stability and other properties of the white noise problem that depend on the infinite time interval, continue to hold away from white noise but not far from it. Moreover, we show that near-optimal control laws for linear systems with state-dependent wide-band disturbances may be designed based on an associated natural limiting system with white noise disturbances. The performance degradation which accompanies use of these suboptimal laws is inversely proportional to the bandwidth of the disturbances. Both control and stability problems are singular in the sense that an order reduction occurs in the design of control laws or the computation of stability criteria when the bandwidth of disturbances approaches infinity.

1. INTRODUCTION

The purpose of this work* is to analyze the stability properties and feedback control of systems that are subjected to external noise disturbances. We consider a family of systems indexed by a parameter $\varepsilon > 0$ such that as $\varepsilon \to 0$ the external disturbances become Gaussian white noise. In treating stability questions we assume that the limiting white noise problem has certain properties - for example, sample stability, recurrence, invariant distributions, etc. - and show that if these properties hold in a sufficiently strong sense, then the corresponding problems for $\varepsilon > 0$ and small also have these properites. Our methods provide estimates of the range of bandwidth (proportional to ε^{-2}) for which such conclusions hold and estimates of other relevant parameters, such as rate of convergence of sample trajectories to zero. Typical results of this type are summarized in section 4.

For linear regulator systems with state dependent wide band noise we show that there are natural limiting white noise systems which have linear optimal feedback laws. When these control laws are used in the $\varepsilon > 0$ systems, a degradation occurs in the performance relative to the optimal cost. Explicit expressions are computed for this degradation, which is proportional to the inverse of the bandwidth (that is, ε^2), both for regulators on finite and infinite time intervals. These results are summarized in section 5.

In section 2 we describe the scaling of the problems and justify the terminology "wide-band disturbances." In section 3 we describe the limiting diffusion system ($\varepsilon = 0$) which approximates the behavior of the system with wide band noise ($\varepsilon > 0$ is small).

The systems and the limiting process (bandwidth of the disburbances approaches infinity as $\varepsilon \to 0$) discussed here exhibit the classical features of singular perturbation problems, including order reduction in the limiting diffusion system and a separation of time scales between the natural time scale t of the system state and the fast scale t/ε^2 of the fluctuatios. Boundary layer effects also appear in the regulator problem on a finite time interval [1].

---

[*] This paper is a summary of portions of the work reported in reference [1] which contains complete hypotheses and proofs.

2. STOCHASTIC SYSTEMS WITH WIDE-BAND NOISE DISTURBANCES

Let $z(t)$, $t \geqslant 0$ be the R^n - valued state of a dynamical system and let $y(t)$ be the R^m - valued state of a disturbance process. Consider the model

$$\frac{dz}{dt} = Az(t) + \varepsilon F(z)t),y(t)) + \varepsilon^2 G(z(t),y(t)) \quad z(0) = x_o \varepsilon R^n \tag{2.1}$$

where A is an n x n matrix and $F(z,y)$, $G(z,y)$ are smooth functions on $R^n \times R^m$. Suppose that $y(t)$ is a stationary, ergodic process and that

$$E\, F(z,y(t)) = 0 \tag{2.2}$$

where E denotes expected value. Because of (2.2) the term $F(z(t),y(t))$ plays the role of a fluctuation of the deterministic system $\dot{z}(t) = Az(t)$. Introducing

$$x(t) = e^{-At} z(t), \; \tilde{F}(x,y,t) = e^{-At} F(e^{At}x,y), \; \tilde{G}(x,y,t) = e^{-At} G(e^{At}x,y), \tag{2.3}$$

we find that

$$\frac{dx}{dt} = \varepsilon \tilde{F}(x(t),y(t),t) + \varepsilon^2 \tilde{G}(x(t),y(t),t), x(0) = x_o \tag{2.4}$$

Furthermore, if we set

$$s = \varepsilon^2 t, \; x^\varepsilon(s) = x(s/\varepsilon^2), \; y^\varepsilon(s) = y(s/\varepsilon^2) \tag{2.5}$$

$$F(x,y,s) = \tilde{F}(x,y,s/\varepsilon^2), \; G(x,y,s) = \tilde{G}(x,y,s/\varepsilon^2), \tag{2.6}$$

then $x^\varepsilon(s)$ satisfies

$$\frac{dx^\varepsilon(s)}{ds} = \frac{1}{\varepsilon} F(x^\varepsilon(s),y^\varepsilon(s),s) + G(x^\varepsilon(s),y^\varepsilon(s),s) \quad x^\varepsilon(0) = x_o \,. \tag{2.7}$$

If all the eigenvalues of the matrix A are distinct and purely imaginary, then $x(t)$ describes the fluctuation of $z(t)$ caused by the disturbance terms in (2.1). If the natural scale of the disturbance process is t, then the effects of the disturbance appear in the state in the slow time scale $s = \varepsilon^2 t$.

We may justify this claim and the use of the terminology "wide band disturbances" as follows.

Let $R(x,t) = EF(x,y(s))F^T(x,y(s + t))$ be the covariance matrix of the local fluctuations and let $S(x,\omega) = \int_{-\infty}^{\infty} R(x,t)\, e^{i\omega t} dt$ be the corresponding power spectral density. Suppose that for all x in R^n, $S(x,\omega) = 0$ for $|\omega| > \omega_o > 0$, so that S is bandlimited. Now, following the steps above, let

$F^\varepsilon(x,y^\varepsilon(t) = F(x,y(t/\varepsilon^2))/\varepsilon$. Then the covariance of F^ε satisfies

$$\begin{aligned} R^\varepsilon(x,t) &= EF^\varepsilon(x,y^\varepsilon(s))\, F^{\varepsilon T}(x,y^\varepsilon(t + s)) \\ &= \varepsilon^{-2} R(x,t/\varepsilon^2) \xrightarrow[\varepsilon \downarrow 0]{} \left(\int_{-\infty}^{\infty} R(x,u)\,du\right) \delta(t) \quad \text{(white noise).} \end{aligned} \tag{2.8}$$

The associated power spectral density is $S^\varepsilon(x,\omega) = S(x,\varepsilon^2\omega)$ which satisfies $S^\varepsilon(x,\omega) = 0$ For $|\omega| > \varepsilon^{-2}\omega_o$ as $\varepsilon \to 0$, the bandwidth of S^ε approaches infinity. In this sense $F^\varepsilon(x,y^\varepsilon(t))$ approaches a white noise process with position (x) dependent statistics.

In this work we are interested in the behavior of (2.4) or (2.7) in the limit as $t \to \infty$ in (2.4) or $s \to \infty$ in (2.7) with ε fixed and $0 < \varepsilon \leq \varepsilon_o$, where ε_o is sufficiently small. In the next section we define a natural limit for the solution $x^\varepsilon(s)$ of (2.7) as $\varepsilon \to 0$

3. THE WHITE NOISE LIMIT

Assume for simplicity that $A = 0$ in (2.1) so that (2.7) is

$$\frac{dx^\varepsilon(s)}{ds} = \frac{1}{\varepsilon} F(x^\varepsilon(s), y^\varepsilon(s)) + G(x^\varepsilon(s), y^\varepsilon(s)), \quad x^\varepsilon(s) = x_o \tag{3.1}$$

Assume that for all x, y and some constant k

$$|F(x,y)| \le k(1 + |x|), \quad |G(x,y)| \le k(1 + |x|), \quad \left|\frac{\partial F}{\partial x}\right| \le k, \quad \left|\frac{\partial G}{\partial x}\right| \le k, \tag{3.2}$$

and that F and G are smooth functions of x and y.

The limiting behavior of (3.1) as $\varepsilon \to 0$ is singular in the sense that an order reduction occurs. To make this feature explicit, assume that $y(t) = y^\varepsilon(\varepsilon^2 t)$ is a stationary, ergodic Markov process with compact state space $S \subset R^m$ and infinitesimal generator Q. Then the pair $(x^\varepsilon(s), y^\varepsilon(s))$ solving (3.1) is a Markov process with state space $R^n \times S$ and infinitesimal generator

$$L^\varepsilon f(x,y) = \frac{1}{\varepsilon^2} Qf(x,y) + \frac{1}{\varepsilon} F^T(x,y) \frac{\partial f(x,y)}{\partial x} + G^T(x,y) \frac{\partial f(x,y)}{\partial x}. \tag{3.3}$$

If $y(t)$ satisfies certain mixing hypotheses (see [1,sec.3,4]), then as $\varepsilon \to 0, x^\varepsilon(s)$ approaches (weakly) a diffusion Markov process $x^o(s)$ with state space R^n and infinitesimal generator

$$Lf(x) = \frac{1}{2} \sum_{i,j=1}^{n} a_{ij}(x) \frac{\partial^2 f(x)}{\partial x_i \partial x_j} + \sum_{i=1}^{n} b_i(x) \frac{\partial f(x)}{\partial x_i} \tag{3.4}$$

where

$$b_j(x) = \lim_{T\to\infty} \frac{1}{T} \int_o^T EG_j(x,y(t))\, dt + \lim_{T\to\infty} \frac{1}{T} \int_o^T \int_o^t \sum_{k=1}^{n} E\, F_k(x,y(t)) \frac{\partial F_j(x,y(s))}{\partial x_k}\, ds\, dt \tag{3.5}$$

$$a_{ij}(x) = \lim_{T\to\infty} \frac{1}{T} \int_o^T \int_o^T EF_i(x,y(t))\, F_j(x,y(s))\, ds\, dt \tag{3.6}$$

See [1,sec. 4] or [2] for proofs of these facts and more details on the technical restrictions required on $y(t)$ and $F(x,y)$ and $G(x,y)$.

It follows that we may regard the limiting diffusion process $x^o(s)$ as the solution of an Itô stochastic differential equation $dx^o(s) = b(x^o(s))ds + \sqrt{2}m(x^o(s))dw(s), x^o(0) = x_o$ where $m(x)\, m^T(x) = a(x)$ and $w(s)$ is a standard Wiener process.

Consequently, the system (3.1) (or (2.1)) consisting of a deterministic part subject to fluctuations exhibits the two dominant features of singularity perturbed systems - a separation of time scales indexed by a parameter ε, and an order reduction as $\varepsilon \to 0$.

In the computations of control algorithms and stability criteria that follow it is useful to have some additional hypotheses on the process $y(t)$. Let $P(t,y,A)$, $t \ge 0$, $y \in S$, be the transition probability of y. Let $\overline{P}(\cdot)$ be an ergodic distribution for y which we assume to be the unique limit of $P(t,y,\cdot)$ as $t \to \infty$. We may restrict y so that the convergence is exponential. In this case the equation

$$-(Qf)(y) = g(y) \tag{3.7}$$

with

$$\int_S \overline{P}(dy)\, g(y) = 0 \tag{3.8}$$

has a bound solution, unique up to an additive constant given by

$$f(y) = -(Q^{-1}g) = \int_S g(z) \left(\int_o^\infty P(t,y,dz) - \overline{P}(dz)\, dt\right) \tag{3.9}$$

We refer to (3.7)-(3.9) by the statement that the Fredholm alternative holds for the process $y(t)$.

Example: The damped harmonic oscillator with random spring constant.

Consider the system

$$\ddot{u}(t) + 2\gamma\varepsilon^2\dot{u}(t) + (\omega^2 + \varepsilon y(t))\, u(t) = 0,\; u(0) = x_o\,,\; \dot{u}(0) = \dot{x}_o \tag{3.10}$$

Here $\omega,\gamma,\varepsilon$ are positive constants and $y(t)$ is the random telegraph process; i.e., the two state Markov process with state space $S = \{-\beta,+\beta\}$ with α^{-1} the mean time between jumps. The infinitesimal generator of $y(t)$ is

$$Qf(y) = \alpha\,[-f(y) + f(-y)]. \tag{3.11}$$

and, on functions $g(y)$ such that $\int \overline{P}(dy)g(y) \equiv \frac{1}{2}[g(+\beta) + g(-\beta)\;] = 0$ we have $-(Q^{-1}g)(y) = \;[g(y) - g(-y)]/4\alpha$ Let

$$z(t) = \frac{1}{2}\begin{bmatrix} u(t) - i\dot{u}(t)/\omega \\ u(t) + i\dot{u}(t)/\omega \end{bmatrix} \tag{3.12}$$

then

$$\dot{z}(t) = (A_o + \varepsilon\tilde{A}_1(t) + \varepsilon^2\tilde{A}_2)\; z(t) \tag{3.13}$$

where

$$A_o = i\omega\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},\quad \tilde{A}_1(t) = \frac{1}{2}\, i\omega^{-1}y(t)\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix},\quad \tilde{A}_2 = \gamma\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \tag{3.14}$$

Defining $s = \varepsilon^2 t$ and $x^\varepsilon(s) = e^{-A_o s/\varepsilon^2}\; z(s/\varepsilon^2)$ transforms (3.13) to

$$\frac{dx^\varepsilon(s)}{ds} = [\frac{1}{\varepsilon}\, A_1^\varepsilon(s) + A_2^\varepsilon(s)]\; x^\varepsilon(s),\; x^\varepsilon(0) = [x_o,\dot{x}_o]^T\;;\; s \geq 0 \tag{3.15}$$

where $y^\varepsilon(s) = y(s/\varepsilon^2)$ and

$$A_1^\varepsilon(s) = \frac{1}{2}\; i\omega^{-1}\; y^\varepsilon(s)e^{-A_o s/\varepsilon^2}\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}e^{A_o s/\varepsilon^2} \tag{3.16}$$

$$A_2{}^\varepsilon(s) = \;\gamma\; e^{-A_o s/\varepsilon^2}\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}e^{A_o s/\varepsilon^2}\;. \tag{3.17}$$

As $\varepsilon \to 0$ the process $x^\varepsilon(s)$ converges to a diffusion Markov process with infinitesimal generator

$$Lf(x) = <(\overline{A}_2 + \overline{A_1\; A_1})x, \frac{\partial f(x)}{\partial x} > + \frac{1}{2}\sum_{i,j=1}^{2}\;\;(\overline{A_1 xx^* A_1^*})\;\;\frac{\partial^2 f(x)}{\partial x_i\,\partial x_j} \tag{3.18}$$

where

$$\overline{A}_2 = -\gamma I \tag{3.19}$$

$$\overline{A_1 A_1} = \frac{1}{4\omega^2}\begin{bmatrix} S^*(2\omega) - S(0) & 0 \\ 0 & S(2\omega) - S(0) \end{bmatrix} \tag{3.20}$$

$$\overline{A_1^* A_1} = \frac{1}{4\omega^2}\begin{bmatrix} S^*(2\omega) + S(0) & 0 \\ 0 & S(2\omega) + S(0) \end{bmatrix} \tag{3.21}$$

Here A^* is the Hermitian transpose of A. Also; $S(\nu)$ is the power spectral of the $y(t)$ process $S(\nu) = \int_o^\infty e^{i\nu t}\; Ey(t)y(0)\,dt = \alpha\beta^2/\;(1 + i\alpha\nu)$

Note that the presence of A_o 0 in (3.13) introduces an additional averaging in the computation of L.

4. STOCHASTIC STABILITY BASED ON THE WIDE-BAND NOISE APPROXIMATION

Consider the system (3.1) which defines $(x^\varepsilon(s), y^\varepsilon(s))$ under the hypotheses placed on $y(t)$ in sections 2 and 3. Assume that $F(x,y)$ and $G(x,y)$ satisfy (3.2) and that for all y in S

$$F(0,y) = 0 \quad , \quad G(0,y) = 0 \tag{4.1}$$

Theorem 4.1. Let $V(x)$ be a smooth, positive definite function on R^n such that $V(x) \to +\infty$ as $|x| \to \infty$, and $V(x)$ behaves like a polynominal in $|x|$ for $|x|$ large. Suppose that

$$LV(x) \leq -\gamma V(x), \quad x \varepsilon R^n \tag{4.2}$$

for some $\gamma > 0$ where L is defined by (3.4). Then there is an $\varepsilon_o > 0$ such that for ε fixed, $0 < \varepsilon \leq \varepsilon_o$, $P_{x,y} \{\lim_{s \uparrow \infty} |x^\varepsilon(s)| = 0\} = 1$. Here $P_{x,y}(\cdot)$ is the conditional probability given $x^\varepsilon(0) = x$, $y^\varepsilon(0) = y$.)

Proof. (summary) Recall that L^ε in (3.2) is the infinitesimal generator of $(x^\varepsilon(s), y^\varepsilon(s)$. Let $V^\varepsilon(x,y) = V(x) + \varepsilon V_1(x,y) + \varepsilon^2 V_2(x,y)$ with V_i, $i = 1,2$, to be chosen. Then

$$\begin{aligned} L^\varepsilon V^\varepsilon(x,y) &= \frac{1}{\varepsilon} (QV_1 + F^T \frac{\partial V}{\partial x}) \\ &+ LV(x) + (QV_2 + F^T \frac{\partial V_1}{\partial x} + G^T \frac{\partial V}{\partial x} - LV(x)) \\ &+ \varepsilon (F^T \frac{\partial V_2}{\partial x} + G^T \frac{\partial V_1}{\partial x}) + \varepsilon^2 (G^T \frac{\partial V_2}{\partial x}) \end{aligned} \tag{4.3}$$

We choose V_1 and V_2 so that

$$QV_1 + F^T \frac{\partial V}{\partial x} = 0, \quad QV_2 + F^T \frac{\partial V_1}{\partial x} + G^T \frac{\partial V}{\partial x} - LV(x) = 0 \tag{4.4}$$

(The assumptions on F, G and V guarantee that the Fredholm Alternative (3.7) - (3.9) applies.) Then $V^\varepsilon(x,y)$ satisfies (for all y)

$$c_1 V(x) \leq V^\varepsilon(x,y) \leq c_2 V(x) \tag{4.5}$$

for some positive constants c_i and $\varepsilon \leq \varepsilon'_o$ small. Let $\hat{\gamma} > 0$ be a constant. Using (4.4) in (4.3) we have

$$\begin{aligned} (L^\varepsilon + \hat{\gamma}) V^\varepsilon(x,y) &= \hat{\gamma} V^\varepsilon(x,y) + LV(x) \\ &+ \varepsilon [F^T \frac{\partial V_2}{\partial x} + G^T \frac{\partial V_1}{\partial x}] + \varepsilon^2 G^T \frac{\partial V_2}{\partial x} \\ &\leq (L + c_2 \hat{\gamma} + \varepsilon'_o c_3) V(x) \end{aligned} \tag{4.6}$$

For some c_3. If we choose $\hat{\gamma}$ and $\varepsilon_o \leq \varepsilon'_o$ so that

$$c_2 \hat{\gamma} + \varepsilon_o c_3 \leq \gamma \tag{4.7}$$

then we find that $(L^\varepsilon + \hat{\gamma}) V^\varepsilon(x,y) \leq 0$. Stability follows from this inequality and standard arguments (See [1]). QED

Remarks. (1) The condition (4.2) guarantees that the limiting diffusion process $x(s) = (\text{weak}) \lim_{\varepsilon \downarrow 0} x^\varepsilon(s)$ is uniformly stochastically asymptotically stable as $s \to \infty$. Thus, as desired, stability of the finite bandwidth system provided that the bandwidth ($\approx\varepsilon^{-2}$) of the disturbances is large .

(2) The step of introducing the correctors V_1 and V_2 from (4.4) to eliminate the singular and 0(1) terms in (4.3) is the key to our method here and in [1].

If the system (3.1) is linear, stability conditions may be given which become necessary and sufficient in the limit $\varepsilon \to 0$. Assume that $F(x,y) = A(y)x$, $G(x,y) = B(y)x$ for some n x n matrix functions A, B continuous on S. The limiting generator L in (3.4) is defined by

$$a_{ij}(x) = \int_{-\infty}^{\infty} \sum_{k,1=\ell}^{n} E[A_{ik}(y(t)A_{j\ell}(y(0)]x_k x_\ell dt \tag{4.8}$$

$$b_j(x) = \sum_{k=1}^{n} E[B_{jk}(y(0)x_k] + \int_{-\infty}^{\infty} \sum_{i,k=1}^{n} E[A_{ik}(y(t))\cdot A_{jk}(y(0)]x_i dt$$

Theorem 4.2. Let $x^\varepsilon(s)$ satisfy

$$\frac{dx^\varepsilon(s)}{ds} = [\frac{1}{\varepsilon} A(y^\varepsilon(s)) + B(y^\varepsilon(s))]\, x^\varepsilon(s), \quad x^\varepsilon(0) = x_o \tag{4.9}$$

and suppose that there exists a smooth function h on S^{n-1}, the unit sphere in R^n, such that if

$$f(x) = \log|x| + H(x/|x|), \tag{4.10}$$

then

$$Lf(x) \leq q < 0 \quad , \quad x \in R^n \tag{4.11}$$

where q is a constant. Then for $0 < \varepsilon \leq \varepsilon_0$, ε fixed and ε_o sufficiently small, $P_{x,y}\{\lim_{s\uparrow\infty} |x^\varepsilon(s)| = 0\} = 1$, $(x,y)\varepsilon R^n \times S$.

Remarks. The condition (4.11) is due to Khasminskii [3] and is necessary and sufficient for the associated limiting diffusion process $x^0(t)$ to be almost surely asymptotically stable. The idea of introducing h is due to Pinsky [4].

Example: Consider the harmonic oscillator described by (3.13). Let

$$x_1^\varepsilon(t) = u(t) \quad , \quad x_2^\varepsilon(t) = \dot{u}(t)/\omega \tag{4.12}$$

and assume that y(t) is the random telegraph process. In polar coordinates

$$x_1 = e^r \cos\theta,\ x_2 = e^r \sin\theta,\ -\infty<r<\infty,\ 0\leq\theta<2\pi$$

the infinitesimal generator L^ε of the $(x_1^\varepsilon(t), x_2^\varepsilon(t), y(t))$ process in the scaling of (2.5) is

$$\begin{aligned} L^\varepsilon f(r,\theta,y) = {} & \alpha \quad [-f(r,\theta,y) + f(r,\theta,-y)] - \omega \frac{\partial f(r,\theta,y)}{\partial\theta} \\ & - \varepsilon\omega^{-1} y[\frac{1}{2} \sin 2\theta \frac{\partial f(r,\theta,y)}{\partial r} + \cos^2 \theta \frac{\partial f(r,\theta,y)}{\partial\theta}] \\ & - \varepsilon^2 2\gamma[\sin^2 \theta \frac{\partial f(r,\theta,y)}{\partial r} + \frac{1}{2} \sin 2\theta \frac{\partial f(r,\theta,y)}{\partial\theta}] \end{aligned} \tag{4.13}$$

Note that the non-zero A_o term in (3.13) results in the extra term $-\omega\partial/\partial\theta$ in L^ε and introduces an additional averaging the analysis.

The generator of the limiting diffusion process to which

$r^{\varepsilon}=\frac{1}{2}\log(x_1^{\varepsilon^2}(t)+x_2^{\varepsilon^2}(t))$ converges as $\varepsilon \to 0$, is given by

$$Lf(r) = \frac{1}{2}\sigma^2 \frac{\partial^2 f(r)}{\partial r^2} + (\sigma^2 - \gamma)\frac{\partial f(r)}{\partial r} \tag{4.14}$$

where

$$\sigma^2 = \frac{\alpha\beta^2}{8\omega^2(\omega^2+\alpha^2)} \tag{4.15}$$

That is, the process r(t) generated by L is a one-demensional Brownian motion with variance parameter σ^2 and drift $\sigma^2-\gamma$.

Choosing $f(r)=r$, we have

$$Lf(r) = \sigma^2 - \gamma = q \tag{4.16}$$

and so the stability condition for the limit problem ($q < 0$) is

$$\sigma^2 < \gamma\ . \tag{4.17}$$

An estimate of ε_0 (from (4.6),(4.8)) is

$$\varepsilon_o = [\alpha\beta^2 - 8\omega^2\gamma(\omega^2+\alpha^2)]/[4\omega(\omega+\alpha)]. \tag{4.18}$$

The condition (4.22) is well-known (see, for example, Stratonovich [5]); its extension to the $\varepsilon > 0$, $0 < \varepsilon \leq \varepsilon_0$, case is the novelty here.

5. OPTIMAL STEADY STATE CONTROL BASED ON THE WIDE-BAND NOISE APPROXIMATION

Let y(t) and $y^{\varepsilon}(t) = y(t/\varepsilon^2)$ be the jump Markov process described in Section 3 and let A(y), B(y) and $\tilde{A}(y)$ be nxn, nxn, and nx1 matrix-valued continuous functions on S. Consider the control problem

$$\begin{aligned}\underset{u\varepsilon U}{\text{minimize}}\ \lim_{T\to\infty} E\ \frac{1}{T}\int_0^t u^T(t)u(t) + x^{\varepsilon T}(t)\, Mx^{\varepsilon}(t)dt\\ = J_*^{\varepsilon}\end{aligned} \tag{5.1}$$

$$\frac{dx\varepsilon(t)}{dt} = \frac{1}{\varepsilon}(A[y^{\varepsilon}(t)]x^{\varepsilon}(t) + \tilde{A}[y^{\varepsilon}(t)]) + B[y^{\varepsilon}(t)]x^{\varepsilon}(t) + Cu(t) \tag{5.2}$$

$$x^{\varepsilon}(0) = x_0\ ;\ t \geqslant 0$$

Here C is an nxm matrix and U is a set of controls to be specified. We assume that E A[y(0)] = 0 and E $\tilde{A}$[y(0)] = 0 and that the process A[y(t)], $\tilde{A}$[y(t)] and B[y(t)] are mutually independent. Also, if $\overline{B}$ = E B[y(0)] and

$$\overline{AA} = \int_{-\infty}^{\infty} E(A[y(t)]\,A[y(0)])dt, \tag{5.3}$$

then the pair $(\overline{B}+\overline{AA}, C)$ is stabilizable[6], and

$$\inf_{W\varepsilon R^{nxn}} 2\left|\int_{-\infty}^{\infty} e^{t(\overline{B}+\overline{AA}-CW)^T}\,\overline{A^TA}\,e^{t(\overline{B}+\overline{AA}-CW)}dt\right|<1 \tag{5.4}$$

where $\overline{A^TA}$ is defined like $\overline{AA}$. Finally, we suppose that $M=M^T>0$ and $\overline{\tilde{A}\tilde{A}^T} > 0$.

We may associate a "limiting diffusion control problem" with (5.1), (5.2). Specifically,

$$\min_{u}\lim_{T\to 0} E\,\frac{1}{T}\int_0^T u^T(t)u(t) + x^T(t)Mx(t)dt \tag{5.5}$$

$$dx(t) \doteq Fx(t)dt + Cu(t)dt + Hx(t)db_1(t) + Kdb_2(t)\ ;\ x(0) = x_0,\ t \geq 0 \tag{5.6}$$

where $b_1(t)$, i=1,2, are independent standard Brownian motions and

$$F = \overline{B} + \overline{AA},\ KK^T = \overline{(\tilde{A}\tilde{A}^T)},\ H^TH = \overline{2A^TA} \tag{5.7}$$

The assumption on $\overline{AA}$, $\bar{B}$, etc. guarantee that (5.5), (5.6) has a unique optimal control law [6]

$$u_0(t) = \phi_0[x(t)] = -C^T P_0 x(t) \tag{5.8}$$

where P_0 is the unique positive definite solution of

$$(\bar{B} + \overline{AA})^T P_0 + P_0(\bar{B} + \overline{AA}) + 2\overline{A^T P_0 A} - P_0 CC^T P_0 + M = 0 \tag{5.9}$$

The optimal cost is

$$\lambda_0 = 2 \text{ trace } (P_0 \overline{\tilde{A}\tilde{A}^T}) \tag{5.10}$$

If we set $u(t) = \phi_0[x^\varepsilon(t)]$ in (5.2), then we discover that there is a maximum value of ε, say $\varepsilon_0 > 0$, such that the solution $(x^\varepsilon(t), y^\varepsilon(t))$has invariant measures with finite second moments. (Asymptotic existence of second moments for $x^\varepsilon(t)$ follows from modification of the proof of Theorem 4.1 [1, sec. 6]. Existence of invariant measures is immediate from this and a result of Beneš [7].) We define the set of admissible control laws to be the set of smooth maps $\phi: R^n \times S \to R^m$ so that the resulting solution of (5.2) with $u(t) = \phi(x^\varepsilon(t), y^\varepsilon(t))$ has invariant measures with finite second moments for all ε fixed, $0 < \varepsilon \leq \varepsilon_0$. Note ϕ in (5.8) is admissible.

By modifying Wonham's basic argument [6], we are able to show that the problem (5.1), (5.2) has an optimal control law in this set and that the corresponding optimal cost J_*^ε satisfies

$$|J_*^\varepsilon - \lambda_0 - \varepsilon^2\lambda_2| = 0(\varepsilon^3) \tag{5.11}$$

where λ_2 depends on the data ($\bar{B}$,$\overline{AA}$,M, etc.) (See [1, Theorem 10.1].) If we use the "limiting diffusion control law", ϕ_0, in the problem (5.1), (5.2) then a degradation in performance occurs. Specifically, the associated suboptimal cost, J_s^ε, of the law ϕ_0 in (5.1), (5.2) satisfies

$$|J_s^\varepsilon - \lambda_0 - \varepsilon^2\gamma_2| = 0(\varepsilon^3) \tag{5.12}$$

where γ_2 is a function of the data.

It is natural to call the difference $D^\varepsilon = J_s^\varepsilon - J_*^\varepsilon$, the "performance degradation." Based on the estimates (5.11) (5.12), we have the following
Theorem 5.1: The performance degradation D^ε resulting from use of the limiting diffusion control law ϕ_0 in the problem (5.1), (5.2) satisfies

$$J_s^\varepsilon - J_*^\varepsilon = D^\varepsilon = \varepsilon^2(\gamma_2 - \lambda_2) + 0(\varepsilon^3) \tag{5.13}$$

where

$$\gamma_2 - \lambda_2 = \text{trace } [(E_2 + P_0 CC^T P_0) \overline{\tilde{A}\tilde{A}^T}] \tag{5.14}$$

and $E_2 > 0$ satisfies the linear equation

$$\hat{B}^T E_2 + E_2\hat{B} + 2\overline{A^T E_2 A} + \overline{(A^T P_0 + P_0 A)CC^T(A^T P_0 + P_0 A)} = 0 \tag{5.15}$$

Here $\hat{B} = \bar{B} + \overline{AA} - CC^T P_0$ and the overbar in (5.15) is the averaging operation (of (5.3), for example).

Note that the cost degradation $\gamma_2 - \lambda_2$ depends only on first and second order statistics of the data. It follows that control designs which minimize this degradation as well as the cost J^ε will be robust in the presense of variations in the underlying process y(t) other than its bandwidth and second order statistics. This aspect of the design problem will be reported elsewhere.

The proof of Theorem 5.1 [1, sec. 10] involves first computing the approximations (5.11) and (5.12). The latter is relatively straightforward, involving expansion of the solution to a steady-state covariance equation. The estimate (5.11) is more difficult. Following Wonham [6], we identify a Hamilton-Jacobi equation

$$\lambda^\varepsilon = \frac{1}{\varepsilon^2} QV^\varepsilon + \frac{1}{\varepsilon}(x^T A^T + \tilde{A}^T)V_x^\varepsilon + x^T B^T V_x^\varepsilon - V_x^{\varepsilon T} CC^T V_x^\varepsilon + X^T Mx. \qquad (5.16)$$

$$= \frac{1}{\varepsilon^2} L_1 V^\varepsilon + \frac{1}{\varepsilon} L_2 V^\varepsilon + L_3(V^\varepsilon) + x^T Mx$$

that is, we must find a constant λ^ε and a smooth $V^\varepsilon(x,y)$ that satisfy (5.16). The value $\lambda^\varepsilon = J_*^\varepsilon$. Equation (5.16) may be "solved" by picking V^ε to be a quadratic form in x. This leaves the equation ($P^\varepsilon = P^\varepsilon(y)$)

$$\frac{1}{\varepsilon^2} QP^\varepsilon + \frac{1}{\varepsilon}(A^T P^\varepsilon + P^\varepsilon A) + B^T P^\varepsilon + P^\varepsilon B - P^\varepsilon CC^T P^\varepsilon + M = 0. \qquad (5.17)$$

to be analyzed. We assume $P^\varepsilon(y) = P_0 + \varepsilon P_1(y) + \varepsilon^2 P_2(y) + \ldots$ and use the Fredholm Alternative to find equations from $P_0, P_1, \ldots$. P_0 satisfies (5.9) and the higher order terms are based on this. These in turn provide the series $\lambda^\varepsilon = \lambda_0 + \varepsilon^2\lambda_2 + \cdots$.

To estimate the error (5.11), we define $V_0(x) = x^T P_0 x$, the solution to the "limiting Hamilton-Jacobi" equation, and define the error $\tilde{Z}^\varepsilon = V^\varepsilon(x,y) - V_0(x)$. The error satisfies a Hamilton-Jacobi type equation with singular terms. To eliminate the singular terms we define the "correctors"

$$\chi_1(x,y) = -(L_1^{-1} L_2 V_0)(x,y)$$

$$\chi_2(x,y) = -[L_1^{-1}(L_2\chi_1 + L_3(V_0) + x^T Mx - \lambda_0)](x,y)$$

and the modified error $Z^\varepsilon = V^\varepsilon - (V_0 + \varepsilon\chi_1 + \varepsilon^2\chi_2)$. Then Z^ε satisfies a Hamilton-Jacobi equation with no singular terms. We may regard Z^ε as the "cost" function for a control problem and obtain the desired estimate from the properties of this problem.

Just as in the stability Theorem 4.1 and the other results in [1] the role of the Fredholm Alternatives and the use of the correctors is essential.

REFERENCES

1. G. Blankenship, G. Papanicolaou, "Stability and Control of Strochastic Systems with Wide-Band Noise Distrubances I, II" submitted for publication.

2. G. Papanicolaou, W. Kohler, "Asymptotic Theory of Mixing Stochastic Ordinary Differential Equations," Comm. Pure and Appl. Math., 27(1974), pp. 641-668.

3. R. Khas'minskii, "Necessary and Sufficient Conditions for the Asymptotic Stability of Linear Stochastic Systems," Theory Probability and Appl., 12(1967), pp. 144-147.

4. M. Pinsky, "Stochastic Stability and the Dirichlet Problem," Comm. Pure and Appl. Math., 27(1974), pp. 311-350.

5. R. Stratonovich, Topics in the Theory of Random Noise, Gordon and Breach, New York, 1963.

6. W. Wonham, "Random Differential Equations in Control Theory," in Probabilistic Methods in Applied Mathematics, A. Bharacha-Reid, ed., Vol. 2, pp. 131-212, Academic Press, New York, 1970.

7. V. Beneš, "Finite Regular Invariant Measures for Feller Processes," J. Appl. Probability, 5(1968), pp. 203-209.

STOCHASTIC CONTROL OF LINEAR SINGULARLY PERTURBED SYSTEMS*

A. H. Haddad and P. V. Kokotovic
Coordinated Science Laboratory and Department of Electrical Engineering
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

ABSTRACT

This paper applies singular perturbation theory to the stochastic control for the Linear-Quadratic-Gaussian (LQG) problem for systems with fast and slow modes. The limiting behavior of the optimal control and the performance index is investigated. It is shown that the optimal control can be approximated by a near optimal control which is obtained as a combination of a slow control and a fast control computed in separate time-scales.

I. INTRODUCTION

The singular perturbation approach [1] to the linear regulator and filtering has decomposed the problem into two lower order problems treated separately in different time scales for fast and slow subsystems [2-9]. As a continuation, this paper presents the results for the stochastic control of linear singularly perturbed systems with quadratic performance index and Gaussian noise. Some of the results of the filtering problem [9] and the regulator problem [3] carry over to the combined LQG control problem, but there are important differences.

The system to be controlled is assumed to have been transformed by a nonsingular transformation [10] into the block diagonal form

$$\dot{x} = A_o x + B_o u + G_o w \tag{1}$$

$$\mu \dot{z} = A_2 z + B_2 u + G_2 w \tag{2}$$

with observations

$$y = C_o x + C_2 z + v \tag{3}$$

where x, z, y, and u are n-, m-, q-, and r-dimensional vectors, and w and v are uncorrelated white Gaussian noise vectors with covariances

$$\begin{aligned} E\{w(t)w'(\tau)\} &= Q(t)\,\delta(t-\tau) \\ E\{v(t)v'(\tau)\} &= R(t)\,\delta(t-\tau). \end{aligned} \tag{4}$$

The parameter $\mu > 0$ represents small time constants and similar physical quantities making (2) the fast subsystem since as $\mu \to 0$ then $\dot{z} \to \pm\infty$. There is no loss in generality in considering systems of the form (1), (2) which have been used in the filtering [9] and the time-optimal [6] problems to simplify the derivations. For A_2 nonsingular, every singularly perturbed linear system can be transformed into the form (1), (2).

The problem is to determine the control u(t) as a functional of past observations $\{y(\tau), t_o \le \tau \le t\}$ to minimize the performance index J

$$J = E\left\{\frac{1}{2}\begin{bmatrix} x \\ z \end{bmatrix}' \begin{bmatrix} \Gamma_1 & \mu\Gamma_{12} \\ \mu\Gamma_{12}' & \mu\Gamma_2 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix}\Bigg|_{t=t_f} + \frac{1}{2}\int_{t_o}^{t_f}\left\{\begin{bmatrix} x \\ z \end{bmatrix}' \begin{bmatrix} L_1 & L_{12} \\ L_{12}' & L_2 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + u'Su\right\}dt\right\} \tag{5}$$

Motivated by previous results for the regulator [3] and filtering [9] problems, the objective of this paper is to decompose the optimal stochastic control problem into two separate problems for slow and fast subsystems, and to investigate the near optimality and limiting behavior of the solutions.

Depending on the stability of the matrix $A_2(t)$, one of the two following assumptions is fundamental:

Assumption 1: For every fixed $t \in [t_o, t_f]$, the eigenvalues of $A_2(t)$ have negative real parts, $\mathrm{Re}\lambda\{A_2(t)\} \leq \delta < 0$.

Assumption 2: For every fixed $t \in [t_o, t_f]$, the matrix $A_2(t)$ is nonsingular the pair $[A_2(t), C_2(t)]$ is detectable, and the pair $[A_2(t), B_2(t)]$ is stabilizable.

In addition it is assumed that all the matrices appearing in (1)-(5) are continuously differentiable functions of t.

In the next section a fast control problem will be considered first. Then the complete system solution will be discussed and a reduced-order problem will formally be defined. Finally, the study of the limiting behavior of the complete system solution will be made by using both the reduced and the fast problems.

II. A FAST CONTROL PROBLEM

The fast control problem considered here may be formulated as having the system state and observation equations

$$\mu\dot{z} = A_2 z + B_2 u + G_2 w \tag{6}$$

$$y_f = C_2 z + v \tag{7}$$

The problem is to determine the control u as a functional of past observations y_f which minimizes the performance index J_f

$$J_f = E\left\{\frac{\mu}{2}[z'\Gamma_2 z]_{t=t_f} + \frac{1}{2}\int_{t_o}^{t_f}[z'L_2 z + u'Su]\,dt\right\} \tag{8}$$

The optimal control u_f is obtained in a straightforward manner by using the separation principle, so that

$$u_f = -S^{-1}B'_2 K_2 \hat{z}_f \tag{9}$$

where K_2 satisfies the Riccati equation

$$\mu\dot{K}_2 = -[A_2'K_2 + K_2A_2 + L_2] + K_2B_2S^{-1}B_2'K_2,\ K_2(t_f) = \Gamma_2 \tag{10}$$

Here $\hat{z}_f$ denotes the optimal estimate of z(t) given the past observations, which for any given u(t) is the output of the filter

$$\mu\dot{\hat{z}}_f = A_2\hat{z}_f + P_2C_2'R^{-1}[y_f - C_2\hat{z}_f] + B_2u,\ \hat{z}_f(t_o) = E\{z(t_o)\} \tag{11}$$

where $P_2(t)/\mu$ is the error covariance of $\hat{z}_f$ satisfying

$$\mu\dot{P}2 = A_2P_2 + P_2A_2' + G_2QG_2' - P_2C_2'R^{-1}C_2P_2,\ P_2(t_o) = \mu\mathrm{cov}(z(t_o)) \tag{12}$$

which does not depend on u(t). The resulting minimum value J_{f_o} of J_f is

$$J_{fo} = \frac{1}{\mu}\int_{t_o}^{t_f}\mathrm{tr.}\left\{\left[\frac{1}{2}L_2 + K_2P_2C_2'R^{-1}C_2\right]P_2\right\}dt + \frac{1}{2}\,\mathrm{tr.}[P_2(t_f)\Gamma_2]$$
$$+ \frac{1}{2}\,\mu\hat{z}(t_o)K_2(t_o)\hat{z}(t_o). \tag{13}$$

If a control u_1^* satisfying

$$u_f^* = u_f + u_1 \equiv -S^{-1}B_2'K_2\,\hat{z}_f + u_1 \tag{14}$$

is applied to (6) as shown in Figure 1, then the value of J_f^* of J_f can be expressed by

$$J_f^* \equiv J_{f_o} + \Delta J = J_{f_o} + \frac{1}{2}\int_{t_o}^{t_f} E\{u_1' S u_1\} dt. \tag{15}$$

These results are now used to establish the limiting behavior of the optimal control and the performance index for the fast system (6).

The properties of $\hat{z}_f$ and P_2 as $\mu \to 0+$ have already been considered in [9] and are briefly summarized here for completeness. If either Assumption 1 or 2 hold, then the estimate of $\hat{z}_f$ and the covariance P_2/μ of the error $(z - \hat{z}_f)$ for any given u are

$$\hat{z}_f = \bar{z}_o + \tilde{z}_o(\theta) + O(\mu^{\frac{1}{2}}), \quad \theta = (t - t_o)/\mu \tag{16}$$

$$P_2 = \bar{P}_2 + \tilde{P}_2(\theta) + O(\mu) \tag{17}$$

where $\bar{P}_2$ is the solution of the algebraic Riccati equation for every fixed $t \in [t_o, t_f]$

$$0 = A_2\bar{P}_2 + \bar{P}_2 A_2' + G_2 Q G_2' - \bar{P}_2 C_2' R^{-1} C_2 \bar{P}_2 \tag{18}$$

and $\tilde{P}_2(\theta)$ satisfies the boundary layer equation at t_o

$$\frac{d}{d\theta}\tilde{P}_2(\theta) = \bar{A}_2(t_o)\tilde{P}_2 + \tilde{P}_2\bar{A}_2'(t_o) - \tilde{P}_2 C_2'(t_o) R^{-1} C_2(t_o)\tilde{P}_2, \quad \theta \geq 0,$$

$$\tilde{P}_2(0) = \mu \operatorname{cov}(z(t_o)) - \bar{P}_2(t_o) \tag{19}$$

$$\bar{A}_2 \equiv [A_2 - \bar{P}_2 C_2' R^{-1} C_2] \tag{20}$$

The estimate $\bar{z}_o$ is the output of a stationary filter in a stretched time scale $\tau = (t-t_1)/\mu$ for $t_1 \leq t \leq t_1 + \epsilon$

$$\frac{d}{d\tau}\bar{z}_o = A_2(t_1)\bar{z}_o + \bar{P}_2(t_1) C_2'(t_1) R^{-1}[y_f - C_2\bar{z}_o] + B_2 u \tag{21}$$

where t_1 is any fixed time instant in $[\epsilon, t_f - \epsilon]$.

The boundary layer term $\tilde{z}_o(\theta)$ satisfies

$$\frac{d}{d\theta}\tilde{z}_o(\theta) = \bar{A}_2(t_o)\tilde{z}_o(\theta) + \tilde{P}_2(\theta) C_2'(t_o) R^{-1}[y_f - C_2\bar{z}_o] \tag{22}$$

and its covariance is bounded by

$$\frac{1}{\mu} M_1 \exp - \alpha \frac{(t-t_o)}{\mu} \tag{23}$$

Hence for $t \in [t_1, t_f]$, where $t_1 > t_o$ we have

$$\hat{z}_f = \bar{z}_o + O(\mu^{\frac{1}{2}}) \tag{24}$$

$$P_2 = \bar{P}_2 + O(\mu) \tag{25}$$

The limiting behavior of the optimal control is obtained when the expression for K_2 is used in conjunction with the filter. The behavior of K_2 has been analyzed in [3] and under Assumptions 1 or 2 it may be written as

$$K_2 = \bar{K}_2 + \tilde{K}_2(\sigma) + O(\mu), \quad \sigma = (t_f - t)/\mu \tag{26}$$

The gain $\bar{K}_2$ is the solution of the algebraic Riccati equation

$$0 = A_2'\bar{K}_2 + \bar{K}_2 A_2 + L_2 - \bar{K}_2 B_2 S^{-1} B_2' \bar{K}_2 \tag{27}$$

and the boundary layer term $\tilde{K}_2(\sigma)$ at t_f satisfies

$$\frac{d}{d\sigma}\tilde{K}_2(\sigma) = \tilde{A}_2'(t_f)\tilde{K}_2 + \tilde{K}_2\tilde{A}_2(t_f) - \tilde{K}_2B_2(t_f)S^{-1}B_2'(t_f)\tilde{K}_2, \quad \sigma \geq 0$$

$$\tilde{K}_2(0) = \Gamma_2 - \bar{K}_2(t_f) \qquad (28)$$

$$\tilde{A}_2 \equiv A_2 - B_2S^{-1}B_2'\bar{K}_2 \qquad (29)$$

The boundary layer term $\tilde{K}_2$ may be bounded by

$$|\tilde{K}_2| \leq M_2 \exp - \frac{\beta(t_f-t)}{\mu} \qquad (30)$$

so that for $t \in [t_o,t_2]$, where $t_2 < t_f$, we may write

$$K_2 = \bar{K}_2 + O(\mu) \qquad (31)$$

In this way the optimal control is approximated by

$$u_f = \bar{u}_f + u_{f\ell}(\theta) + u_{fr}(\sigma) + O(\mu^{\frac{1}{2}}), \quad t \in [t_o,t_f] \qquad (32)$$

where

$$\bar{u}_f = - S^{-1}B_2'\bar{K}_2\bar{z}_o \qquad (33)$$

and $u_{f\ell}$ and u_{fr} are the layer correction terms

$$u_{f\ell}(\theta) = - S^{-1}B_2'\bar{K}_2\tilde{z}_o(\theta), \quad u_{fr}(\sigma) = - S^{-1}B_2'\tilde{K}_2(\sigma)\bar{z}_o \qquad (34)$$

If these terms are ignored the approximation

$$u_f = \bar{u}_f + O(\mu^{\frac{1}{2}}) \qquad (35)$$

is valid on a subinterval $[t_1,t_2] \subset [t_o,t_f]$.

In analyzing the effect of replacing u_f by its approximations (32) of (35) several cases are of interest. It is easy to verify that

$$E\{u_{f\ell}'Su_{f\ell}\} \leq \frac{1}{\mu}\,\mathrm{tr.}[\bar{K}_2(t_o)B_2S^{-1}B_2'\bar{K}_2(t_o)]M_1 \exp - \frac{\alpha(t-t_o)}{\mu} \qquad (36)$$

so that if $\bar{K}_2(t_o) = O(\mu^i)$ then

$$J_{f\ell} \equiv \frac{1}{2}\int_{t_o}^{t_f} E\{u_{f\ell}'Su_{f\ell}\}dt = O(\mu^{2i}) \qquad (37)$$

Similarly,

$$E\{u_{fr}'Su_{fr}\} \leq \frac{1}{\mu}\,\mathrm{tr.}[M_2B_2S^{-1}B_2'M_2]M_3 \exp - \frac{\beta(t_f-t)}{\mu} \qquad (38)$$

and if $M_2 = O(\mu^j)$, then we have

$$J_{fr} \equiv \frac{1}{2}\int_{t_o}^{t_f} E\{u_{fr}'Su_{fr}\}dt = O(\mu^{2j}) \qquad (39)$$

To see the effect of replacing the optimal control u_f by the control $u_f^* = \bar{u}_f$ we use (14), (15) with

$$u_1 = - [u_{f\ell} + u_{fr}] + O(\mu^{\frac{1}{2}})$$

which, when combined with (37), (39) results in the performance index $\bar{J}_f$

$$\bar{J}_f = J_{fo} + J_{f\ell} + J_{fr} + O(\mu) \qquad (40)$$

Two cases need to be considered to determine the relative magnitude of J_{fo}, $J_{f\ell}$, and J_{fr}. It can easily be seen from (13) that in general $J_{fo} = O(1/\mu)$ which is due to the white noise behavior of the fast variables in the limit as $\mu \to 0+$. If now $L_2 \neq 0$ in (27) as $\mu \to 0+$, then $J_{f\ell} = O(1)$, $J_{fr} = O(1)$, so that the relative error in performance is expressed by

$$\bar{J}_f = J_{fo}[1 + O(\mu)] + O(\mu), \quad J_{fo} = O(\tfrac{1}{\mu}) \tag{41}$$

If now $L_2 = \mu\bar{L}_2$ in (27), then $\bar{K}_2 = O(\mu)$ provided A_2 is stable, otherwise $\bar{K}_2$ will tend to a nonzero limit. In this case from (28)

$$J_{fo} = O(1), \qquad J_{f\ell} = O(\mu^2), \qquad J_{fr} = O(\|\Gamma_2\|^2) \tag{42}$$

Consequently, if in addition we require $\Gamma_2 = \mu^{\frac{1}{2}}\bar{\Gamma}_2$ then $J_{fr} = O(\mu)$, and (41) is still valid. It should be noted that while in both cases the relative increments in the performance index resulting from neglecting the boundary layers are $O(\mu)$, the absolute increments are $O(1)$ in the first case and $O(\mu)$ in the second. The importance of the latter case is also evident from the fact that $\bar{u}_f = O(\mu^{\frac{1}{2}})$ so that since the system is stable no control is really necessary due to the fact that $L_2 = O(\mu)$. [The fact that $J_{fo} = O(1)$ stems from the covariance of the fast states which behave like white noise as $\mu \to 0+$]. In the case of finite L_2, in order to achieve an absolute $O(\mu)$ approximation to J_{fo}, the near optimal control should include the sum of $\bar{u}_f$ and the two layer correction terms.

These results are summarized in the following

Theorem 1: Let Assumption 2 hold, then the limit as $\mu \to 0+$ of the solution of the fast optimal control problem (6)-(8) satisfies (32). Furthermore, the limits of the performance index when u_f or $\bar{u}_f$ are used are given by (40) and (41). If instead Assumption 1 is satisfied and $L_2 = \mu\bar{L}_2$, $\Gamma_2 = \mu^{\frac{1}{2}}\bar{\Gamma}_2$, then (42) holds.

III. THE COMPLETE CONTROL PROBLEM

We now return to the stochastic control problem for the complete system (1)-(5). The solution when $\mu > 0$ is

$$u_o = -S^{-1}\{[B_o'K_1 + B_2'K_{12}']\hat{x}_o + [\mu B_o'K_{12} + B_2'K_2]\hat{z}_o\} \tag{43}$$

where the gains satisfy the Riccati system

$$\dot{K}1 = -(A_o'K_1 + K_1A_o + L_1) + (K_1B_o + K_{12}B_2)S^{-1}(K_1B_o + K_{12}B_2)', \quad K_1(t_f) = \Gamma_1 \tag{44}$$

$$\mu\dot{K}_{12} = -(K_{12}A_2 + \mu A_o'K_{12} + L_{12}) + (K_1B_o + K_{12}B_2)S^{-1}(\mu B_o'K_{12} + B_2'K_2), \quad K_{12}(t_f) = \Gamma_{12} \tag{45}$$

$$\mu\dot{K}_2 = -(A_2'K_2 + K_2A_2 + L_2) + (K_2B_2 + \mu K_{12}'B_o)S^{-1}(\mu B_o'K_{12} + B_2'K_2), \quad K_2(t_f) = \Gamma_2 \tag{46}$$

Similarly, the optimal estimates $\hat{x}_o$ and $\hat{z}_o$ are obtained from the linear filters which for any given control u are

$$\dot{\hat{x}}_o = A_o\hat{x}_o + (P_1C_o' + P_{12}C_2')R^{-1}\,\nu + B_o u \tag{47}$$

$$\mu\dot{\hat{z}}_o = A_2\hat{z}_o + (\mu P_{12}'C_o' + P_2C_2')R^{-1}\,\nu + B_2 u \tag{48}$$

where $\nu(t)$ is the innovation

$$\nu(t) = y(t) - C_o\hat{x}_o(t) - C_2\hat{z}_o(t) \tag{49}$$

The error covariances satisfy

$$\dot{P}_1 = A_oP_1 + P_1A_o' + G_oQG_o' - (P_1C_o' + P_{12}C_2')R^{-1}(C_oP_1 + C_2P_{12}'), \tag{50}$$

$$P_1(t_o) = \text{Cov}\,[x(t_o)]$$

$$\mu\dot{P}_{12} = \mu A_oP_{12} + P_{12}A_2' + G_oQG_2' - (P_1C_o' + P_{12}C_2')R^{-1}(\mu C_oP_{12} + C_2P_2) \tag{51}$$

$$P_{12}(t_o) = \text{Cov}\,[x(t_o),z(t_o)]$$

$$\mu\dot{P}_2 = A_2P_2 + P_2A_2' + G_2QG_2' - (\mu P_{12}'C_o' + P_2C_2')R^{-1}(\mu C_oP_{12} + C_2P_2), \tag{52}$$

$$P_2(t_o) = \mu\,\text{Cov}\,[z(t_o)]$$

To investigate the limiting behavior of the optimal control, we analyze the limiting behavior of the filter (47), (48) and of the linear regulator (43). We first note that expressions analgous to (14) and (15) are valid for J* obtained with u* and J_o obtained with u_o. The dominant contribution to J_o is due to the fast subsystem since it contains an $O(1/\mu)$ term. Therefore, if the slow subsystem is of interest, any approximation or the limiting behavior should include both the $O(1/\mu)$ and the $O(1)$ terms. Under the conditions of Theorem 1, the contributions of the fast subsystem may be reduced to $O(1)$, comparable to the contribution of the slow subsystem.

The first step in a singular perturbation analysis is to define a reduced problem. The reduced problem is formally obtained from (1)-(5) by setting $\mu = 0$. We point out that this is not a valid limiting operation. If A_2 is stable, it is valid for the purpose of substituting z in a slow system (Theorem 1 in [9]). However, it is not valid in general for the purpose of substituting z into (5). The substitution of $N = -C_2A_2^{-1}B_2$, $D = -C_2A_2^{-1}G_2$ and

$$z_r = -A_2^{-1}[B_2u + G_2w] \tag{53}$$

in (1), (3) and (5) results in

$$\dot{x}_r = A_ox_r + B_ou + G_ow \tag{54}$$

$$y = C_ox_r + Nu + Dw + v \tag{55}$$

$$J = E\{\tfrac{1}{2}\int_{t_o}^{t_f}[x_r'L_1x_r - 2x_r'L_{12}A_2^{-1}B_2u + u'S_ou]dt + \tfrac{1}{2}[x_r'\Gamma_1x_r]\Big|_{t=t_f}\} + J_1 \tag{56}$$

$$S_o = S + B_2'A_2'^{-1}L_2A_2^{-1}B_2 \tag{57}$$

Here J_1 contains terms not influenced by the control, some of which, such as the integral of the variance of the white noise w, are ill-defined. Therefore, only the terms which may be affected by the optimization are given in (56). The optimal control u_r which minimizes (56) for the system (54), (55) is called the reduced control and is given by

$$u_r = -S_o^{-1}[K_oB_o - L_{12}A_2^{-1}B_2]'\hat{x}_r \tag{58}$$

where the reduced filter for any given u satisfies

$$\dot{\hat{x}}_r = A_o\hat{x}_r + [P_oC_o' + G_oQD']\,R_o^{-1}[y - C_o\hat{x}_r - Nu] + B_ou, \tag{59}$$

$$R_o = (R + DQD').$$

The gain K_o and error covariance P_o satisfy the usual Riccati equations

$$\dot{K}_o = -(K_oA_o + A_o'K_o + L_1) + (K_oB_o - L_{12}A_2^{-1}B_2)S_o^{-1}(K_oB_o - L_{12}A_2^{-1}B_2)' \tag{60}$$

$$\dot{P}_o = A_oP_o + P_oA_o' + B_oQB_o' - (P_oC_o' + G_oQD')R_o^{-1}(C_oP_o + DQG_o'). \tag{61}$$

The asymptotic validity of the reduced problem is established in two steps. First, the limiting behavior of the filters (47), (48) is investigated for any given u. It follows from Theorem 2 of [9] that under Assumptions 1 or 2 for $t \in [t_1, t_f]$, $t_1 > t_o$ that

$$\begin{gathered} \hat{x}_o = \hat{x}_r + 0(\mu^{\frac{1}{2}}) \\ P_1 = P_o + 0(\mu), \qquad P_2 = \bar{P}_2 + 0(\mu) \end{gathered} \tag{62}$$

Similarly, using [3,11] for $t \in [t_o, t_2]$, $t_2 < t_f$, we have

$$\begin{gathered} K_1 = K_o + 0(\mu), \quad K_2 = \bar{K}_2 + 0(\mu) \\ K_{12} = -(L_{12} - K_oB_oS^{-1}B_2'\bar{K}_2)\tilde{A}_2^{-1} + 0(\mu) \equiv \bar{K}_{12} + 0(\mu) \end{gathered} \tag{63}$$

Now all that we need to obtain the behavior of u_o as $\mu \rightarrow 0+$ is to substitute u_o into (48) and then find the limiting behavior of $\hat{z}_o$. The substitution of (43) into (48) yields

$$\mu\dot{\hat{z}}_o = \tilde{A}_2\hat{z}_o + P_2C_2'R^{-1}\nu(t) - B_2S^{-1}(B_o'K_o + B_2'\bar{K}_{12}')\hat{x}_o + 0(\mu^{\frac{1}{2}}) \tag{64}$$

The usual method in singular perturbations for the separation of fast and slow modes may be used to show that

$$\hat{z}_o = \tilde{A}_2^{-1}B_2S^{-1}(B_o'K_o + B_2'\bar{K}_{12}')\hat{x}_o + \hat{z}_f + 0(\mu^{\frac{1}{2}}) \tag{65}$$

where

$$\mu\dot{\hat{z}}_f = A_2\hat{z}_f + P_2C_2'R^{-1}\nu(t) - B_2S^{-1}B_2'K_2\hat{z}_f \tag{66}$$

The term $\hat{x}_o$ in (65) is only needed insofar as its effect on the slow system is concerned, and its relative contribution to the variance of $\hat{z}_o$ is $0(\mu)$. The second term $\hat{z}_f$ is clearly the state of a fast system with white noise input $\nu(t)$. The substitution of (65) in (43) results after derivations similar to [11] in the expression

$$\begin{aligned} u_o &= -S_o^{-1}(K_oB_o - L_{12}A_2^{-1}B_2)'\hat{x}_o - S^{-1}B_2'K_2\hat{z}_f + 0(\mu^{\frac{1}{2}}) \\ &= -S^{-1}(K_oB_o - L_{12}A_2^{-1}B_2)'\hat{x}_r - S^{-1}B_2'K_2\hat{z}_f + 0(\mu^{\frac{1}{2}}) \\ &\equiv u_r + u_f + 0(\mu^{\frac{1}{2}}) \end{aligned} \tag{67}$$

The definition of u_f in (67) together with $\hat{z}_f$ in (66) imply that u_f and $\hat{z}_f$ are the same as those discussed in Section II if we let $y_f = (y - C_o\hat{x}_r)$. It is seen that the optimal control of the complete system is given approximately as a sum of the reduced control obtained formally and the fast control which is performed in a stretched time scale and is composed of a stationary term, and two boundary layers.

IV. NEAR OPTIMUM PERFORMANCE

The effect of using u_r or $u_r + \bar{u}_f$ in the singularly perturbed system (1), (2) on the performance index will now be considered. Again, it is sufficient to use (14), (15) and (32). The effect of using only the reduced control u_r is

$$J_r = J_o + J_{f\ell} + J_{fr} + \frac{1}{2\mu}\int_{t_o}^{t_f} \mathrm{tr.}[\bar{K}_2 B_2 S^{-1} B_2' \bar{K}_2(\bar{V}_2 - \bar{P}_2)]\,dt + 0(\mu) \tag{68}$$

where $\bar{V}_2/\mu$ is the covariance of z. It is seen therefore that the error $(J_r - J_o)$ is $0(1/\mu)$ unless the fast system is stable and $L_2 = \mu\bar{L}_2$. In such a case neglecting the fast control causes an error of 0(1). If in addition $\Gamma_2 = \mu^{\frac{1}{2}}\bar{\Gamma}_2$, that is the fast variables are stable and are of little interest, then avoiding the use of the fast control results in $0(\mu)$ approximation to the optimal performance. On the other hand, if the fast subsystem is unstable or if it is of sufficient interest, then using a sum of the reduced control and the fast control results in $0(\mu)$ approximation to the optimal performance.

These results for the case when the fast subsystem is asymptotically stable are summarized as follows.
Theorem 2: Let Assumption 1 hold, and let $L_2 = 0(\mu)$ and $\Gamma_2 = 0(\mu^{\frac{1}{2}})$, then the solution to the optimal control problem (1)-(5) as $\mu \to 0+$ is approximated by

$$u_o = u_r + 0(\mu^{\frac{1}{2}}), \qquad t \in [t_1,t_2] \subset [t_o,t_f].$$

where u_r is the optimal control of the reduced problem. If u_r is applied to the system (1), (2), the resulting performance index J_r is near optimal in the sense that

$$J_r = J_o + 0(\mu). \tag{69}$$

A more general case is summarized in the following theorem.
Theorem 3: If Assumption 2 holds then the optimal control u_o for the system (1)-(5) as $\mu \to 0+$ is approximated by

$$\begin{aligned} u_o = u_r + u_f + 0(\mu^{\frac{1}{2}}) &= u_r + \bar{u}_f + u_{f\ell} + u_{fr} + 0(\mu^{\frac{1}{2}}) \\ &\equiv u^* + 0(\mu^{\frac{1}{2}}) \end{aligned} \tag{70}$$

If u^* is applied to the system (1), (2) the resulting performance index J^* is near optimal in the sense that

$$J^* = J_o + 0(\mu). \tag{71}$$

We note that in this case J_o may be $0(1/\mu)$.
The significance of Theorem 3 is that the separation into two time-scales solution is valid for both the filter and the controller as was the case for each separately. The result may be illustrated in Figure 2. While the input to the slow filter incorporates the entire control $u_r + u_f$, which includes a fast component, it is possible without loss of information to use a low pass filter so that such inputs can be sampled at the low rate used for the slow filter. This leads to important simplifications in real time operation where the slow computations are performed by separate dedicated computers.

ACKNOWLEDGMENT

This work was supported in part by the U.S. Army Research Office under Contract DAAG29-76-G-0154, in part by the Energy Research and Development Administration under Contract U.S. ERDA E(49-18)-2088 and in part by the U.S. Air Force under Grant AFOSR-73-2570.

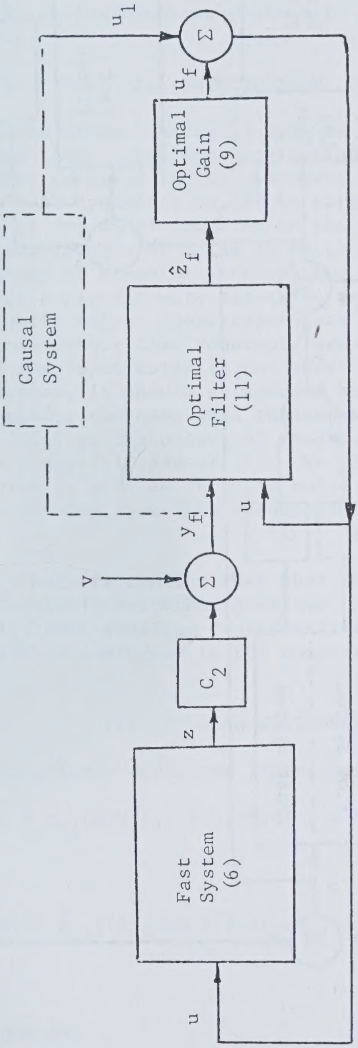

Fast System (6)
u
z
C_2
Σ
v
y_f
Causal System
Optimal Filter (11)
u
$\hat{z}_f$
Optimal Gain (9)
u_f
Σ
u_1

Fig. 1

The definition of u_f, $\hat{z}_f$, and u_1 for the fast system.

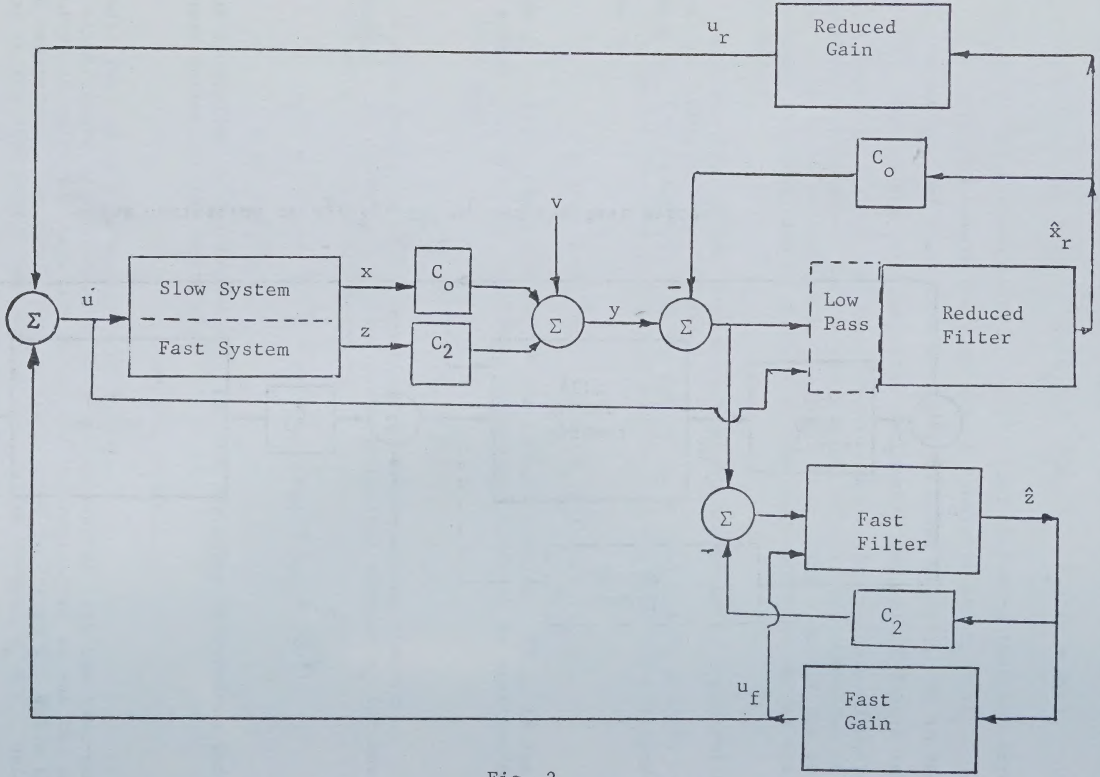

u_r
Reduced Gain
C_o
$\hat{x}_r$
v
Slow System
Fast System
x
z
C_o
C_2
Σ
u
y
$-$
Low Pass
Reduced Filter
Fast Filter
$\hat{z}$
C_2
u_f
Fast Gain

Fig. 2

The near optimal control for the complete system.

ON LINEAR SINGULARLY PERTURBED SYSTEMS WITH STOCHASTIC INPUTS

HASSAN K. KHALIL
Coordinated Science Laboratory
University of Illinois
Urbana, Illinois 61801

ABSTRACT

This paper considers the stochastic control of linear-quadratic problems for singularly perturbed systems when the input noise is colored. A near optimal linear output feedback control is obtained by optimizing a slow subsystem only.

I. INTRODUCTION

An approach to the linear quadratic problem for singularly perturbed systems when both the input noise and observation noise processes are modeled as white noise is given in [5]. However, singularly perturbed systems incorporate fast dynamics; so if the correlation time of the system disturbances is not short compared to the small times of interest of the fast modes, modeling input noise as white is not appropriate, and does not take advantage of singular perturbation techniques. As an illustration consider a circuit with parasitic capacitors. If we model the input noise as white noise, these capacitors are not negligible regardless of how small their time constants are. Thus whenever the correlation time of the input noise is not short compared to times of fast modes of the system, it should be modeled as colored noise. Hence in this paper we consider the case when the input noise is colored. Since colored noise is modeled as the output of a system driven by white noise [8], our problem is a special case of [5]. We will keep the assumption that observation noise is modeled as white noise. The derivation of a control algorithm will show that this assumption does not degrade the results.

The methodology here is simpler than that in previous treatments of linear-quadratic singularly perturbed problems. We approximate a performance index directly, thus avoiding consideration of boundary layers. Moreover the solution is analogous to the compact form of the deterministic case [2].

II. PROBLEM STATEMENT

Consider the singularly perturbed linear system

$$\dot{x}_1(t) = \tilde{A}_{11}(t)x_1(t) + \tilde{A}_{12}(t)x_2(t) + \tilde{B}_1(t)u(t) + \tilde{G}_1(t)v_1(t) \quad , \; x_1(t_o) = x_{10} \tag{1}$$

$$\mu\dot{x}_2(t) = \tilde{A}_{21}(t)x_1(t) + \tilde{A}_{22}(t)x_2(t) + \tilde{B}_2(t)u(t) + \tilde{G}_2(t)v_1(t) \quad , \; x_2(t_o) = x_{20} \tag{2}$$

with observation equation

$$y(t) = \tilde{c}_1(t)x_1(t) + \tilde{c}_2(t)x_2(t) + w_2(t) \tag{3}$$

and performance index:

$$J = E\left\{\frac{1}{2}\begin{pmatrix}x_1\\x_2\end{pmatrix}'\begin{pmatrix}\tilde{T}_1 & \mu\tilde{T}_{12}\\ \mu\tilde{T}'_{12} & \mu\tilde{T}_2\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}\Bigg|_{t=t_f} + \frac{1}{2}\int_{t_o}^{t_f}\begin{pmatrix}x_1\\x_2\end{pmatrix}'\begin{pmatrix}\tilde{c}'_1\tilde{c}_1 & \tilde{c}'_1\tilde{c}_2\\ \tilde{c}'_2\tilde{c}_1 & \tilde{c}'_2\tilde{c}_2\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix} + u'Ru\, dt\right\} \tag{4}$$

where dimensions of x_1, x_2, u, y and v_1 are n, m, r, l and s respectively, μ is a small positive scalar, and $v_1(t)$ is colored noise modeled by

$$\dot{v}_1(t) = H(t)v_1(t) + w_1(t) \qquad , v_1(t_o) = v_{10} \tag{5}$$

$w_1(t)$ and $w_2(t)$ are zero mean white noise processes that in general need not be uncorrelated. The covariance matrix of $w'(t) = (w'_1(t), w'_2(t))$ is

$$E\{w(t)w'(\tau)\} = \begin{pmatrix}V_1(t) & V_{12}(t)\\ V'_{12}(t) & V_2(t)\end{pmatrix}\delta(t-\tau) . \tag{6}$$

Also, x_{10}, x_{20} and v_{10} are random vectors with mean values $\overline{x}_{10}$, $\overline{x}_{20}$ and $\overline{v}_{10}$ and covariance matrices $P_{x_{10}}$, $P_{x_{20}}$ and $P_{v_{10}}$ respectively. They are uncorrelated with $\{w(t), t \geq t_o\}$, and each of x_{10} and x_{20} is uncorrelated with v_{10}.

The problem is to determine the control u(t) as a function of past observation $\{y(\tau), t_o \leq \tau < t\}$ which minimizes (4) under the following conditions for $t\varepsilon[t_o,t_f]$:

1) All the matrices are continuous, bounded and have bounded first derivatives.

2) $\tilde{T}_1 \geq o$, $R(t) > o$, $V_2(t) > o$.

3) $\tilde{A}_{22}(t)$ is a stable matrix $(\mathrm{Re}\{\lambda(\tilde{A}_{22}(t)\} \leq -\sigma$ for some $\sigma > o)$.
   Combining (5) with (1), (2), (3), and (4), and putting $x' = (x'_1, v'_1)$ and $z = x_2$, we obtain the augmented system

$$\dot{x}(t) = A_{11}(t)x(t) + A_{12}(t)z(t) + B_1(t)u(t) + G_1(t)w_1(t) \qquad , x(t_o) = x_o \tag{7}$$

$$\mu\dot{z}(t) = A_{21}(t)x(t) + A_{12}(t)z(t) + B_2(t)u(t) \qquad , z(t_o) = z_o \tag{8}$$

$$y(t) = c_1(t)x(t) + c_2(t)z(t) + w_2(t) \tag{9}$$

with performance index

$$J = E\left\{\frac{1}{2}\begin{pmatrix} x \\ z \end{pmatrix}'\begin{pmatrix} T_1 & \mu T_{12} \\ \mu T'_{12} & \mu T_2 \end{pmatrix}\begin{pmatrix} x \\ z \end{pmatrix}\Bigg|_{t=t_f} + \frac{1}{2}\int_{t_o}^{t_f}\begin{pmatrix} x \\ z \end{pmatrix}'\begin{pmatrix} c'_1c_1 & c'_1c_2 \\ c'_2c_1 & c'_2c_2 \end{pmatrix}\begin{pmatrix} x \\ z \end{pmatrix} + u'Ru\ dt\right\} . \tag{10}$$

We define the reduced problem, formally obtained by setting μ=o in (7)-(10), by

$$\dot{x}_r(t) = A_o(t)x_r(t) + B_o(t)u(t) + G_1(t)w_1(t) \qquad , x_r(t_o) = x_o \tag{11}$$

$$y(t) = c_o(t)x_r(t) + D_o(t)u(t) + w_2(t) \tag{12}$$

with performance index

$$J_r = E\{\frac{1}{2} x'_r(t_f)T_1x_r(t_f) + \frac{1}{2}\int_{t_o}^{t_f}[x'_rc'_oc_ox_r + 2u'D'_oc_ox_r + u'(R + D'_oD_o)u]dt\} \tag{13}$$

where

$$A_o = A_{11} - A_{12}A_{22}^{-1}A_{21} \quad , \qquad B_o = B_1 - A_{12}A_{22}^{-1}B_2$$

$$C_o = C_1 - C_2A_{22}^{-1}A_{21} \quad , \qquad D_o = -C_2A_{22}^{-1}B_2 \quad .$$

III. NEAR OPTIMAL LINEAR OUTPUT FEEDBACK CONTROL

A transformation due to Chang [1], which separates the fast and slow modes, is

$$\begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} I_1-\mu ML & -\mu M \\ L & I_2 \end{pmatrix}\begin{pmatrix} x \\ z \end{pmatrix} \tag{14}$$

where M(t) and L(t) matrices satisfy

$$\mu\dot{L} = A_{22}L - A_{21} - \mu L(A_{11} - A_{12}L) \tag{15}$$

$$\mu\dot{M} = -M(A_{22} + \mu LA_{12}) + A_{12} + \mu(A_{11} - A_{12}L)M \quad . \tag{16}$$

Since the initial conditions for L(t) and M(t) are arbitrary, we choose them to be

$$L(t_o) = A_{22}^{-1}(t_o)\ A_{21}(t_o) \tag{17}$$

$$M(t_o) = A_{12}(t_o)\ A_{22}^{-1}(t_o) \tag{18}$$

Applying the usual singular perturbation techniques [6], [9], we obtain for all $t \varepsilon [t_o, t_f]$.

$$L(t) = A_{22}^{-1}(t)\, A_{21}(t) + 0(\mu) \tag{19}$$

$$M(t) = A_{12}(t)\, A_{22}^{-1}(t) + 0(\mu) \quad . \tag{20}$$

Then (7), (8) and (9) becomes

$$\dot{p}(t) = (A_o(t) + 0(\mu))p(t) + (B_o(t) + 0(\mu))u(t) + (G_1(t) + 0(\mu))w_1(t) \;, \qquad p(t_o) = p_o \tag{21}$$

$$\mu\dot{q}(t) = (A_{22}(t) + 0(\mu))q(t) + (B_2(t) + 0(\mu))u(t) + \mu G_2(t)\, w_1(t) \;, \qquad q(t_o) = q_o \tag{22}$$

$$y(t) = (C_o(t) + 0(\mu))p(t) + (C_2(t) + 0(\mu))q(t) + w_2(t) \; . \tag{23}$$

Letting $\eta(t)$ and $\rho(t)$ satisfy the equations

$$\dot{\eta}(t) = A_o(t)\eta(t) + B_o(t)u(t) + G_1(t)w_1(t) \qquad , \quad \eta(t_o) = \eta_o = x_o \tag{24}$$

$$\mu\dot{\rho}(t) = A_{22}(t)\rho(t) + B_2(t)u(t) + \mu G_2(t)w_1(t) \qquad , \quad \rho(t_o) = \rho_o \quad . \tag{25}$$

it can be easily shown that

$$\eta(t) = p(t) + 0(\mu) \tag{26}$$

$$\rho(t) = q(t) + 0(\mu) \quad . \tag{27}$$

Since $A_{22}(t)$ is a stable matrix for all $t \varepsilon [t_o, t_f]$, using lemma 2 of [4] it follows that $\rho(t)$ is bounded as $\mu \to 0$. Then from (14), (26) and (27) we have

$$x(t) = \eta(t) + 0(\mu) \tag{28}$$

$$z(t) = -A_{22}^{-1}(t)\, A_{21}(t)\, \eta(t) + \rho(t) + 0(\mu) \quad . \tag{29}$$

Lemma 1: As an input to slow linear systems, $\rho(t)$ can be approximated to $0(\mu)$ by $-A_{22}^{-1}(t)\, B_2(t)\, u(t)$, for all $t \varepsilon [t_o, t_f]$, that is

$$\int_{t_o}^{t} \rho(t)dt = \int_{t_o}^{t} [-A_{22}^{-1}(t)\ B_2(t)\ u(t)]\ dt + 0(\mu)\ , \qquad t\epsilon[t_o+\epsilon,t_f] \qquad (30)$$

for some fixed $\epsilon > 0$

Proof: From (25) and lemma 2 of [4], we have

$$\rho(t) = \exp[A_{22}(t_o)(\frac{t-t_o}{\mu})]\rho_o + \frac{1}{\mu}\int_{t_o}^{t} \exp[A_{22}(\sigma)(\frac{t-\sigma}{\mu})][B_2(\sigma)u(\sigma) + \mu G_2(\sigma)w_1(\sigma)]\ d\sigma + 0(\mu)\ . \qquad (31)$$

It can be shown, by reversing the order of integration, that

$$\begin{aligned}\int_{t_o}^{t} e(\tau)d\tau &= \mu\{\exp[A_{22}(t_o)(\frac{t-t_o}{\mu})] - I\}\ A_{22}^{-1}(t_o)\rho_o \\ &\quad + \mu\int_{t_o}^{t} \{\exp[A_{22}(\sigma)(\frac{t-\sigma}{\mu})] - I\}\ G_2(\sigma)w_1(\sigma)d\sigma \\ &\quad + \int_{t_o}^{t} \exp[A_{22}(\sigma)(\frac{t-\sigma}{\mu})]\ A_{22}^{-1}(\sigma)B_2(\sigma)u(\sigma)d\sigma \\ &\quad - \int_{t_o}^{t} A_{22}^{-1}(\sigma)B_2(\sigma)u(\sigma)d\sigma + 0(\mu) \qquad (32) \\ &= -\int_{t_o}^{t} A_{22}^{-1}(t)B_2(t)u(t)dt + 0(\mu)\quad , \text{ for all } t\epsilon[t_o+\epsilon,t_f]\ .\end{aligned}$$

Lemma 2:

$$J = J_S + 0(\mu) \qquad (33)$$

where

$$J_S = E\{\frac{1}{2}\ \eta'(t_f)T_1\ \eta(t_f) + \frac{1}{2}\int_{t_o}^{t_f}[\eta'C_o'C_o\eta + 2u'D_o'C_o\eta + u'(R+D_o'D_o)u]dt\}\ . \qquad (34)$$

Proof: It follows from (29) and (30), that for an $0(\mu)$ approximation of J the quantity

$$A_{22}^{-1}(t)B_2(t)\eta(t) - A_{22}^{-1}(t)B_2(t)u(t) \qquad (35)$$

can be used instead of z in (10). Then, using (28) and (35), (10) reduces to

$$J = J_S + 0(\mu)$$

We define

$$y_1(t) = C_o(t)\eta(t) + D_o(t)u(t) + w_2(t) \quad . \tag{36}$$

Lemma 3: As an input to slow linear systems y(t) can be approximated to $O(\mu)$ by $y_1(t)$ for all $t\epsilon[t_o,t_f]$, that is

$$\int_{t_o}^{t} y(t)dt = \int_{t_o}^{t} y_1(t)dt + O(\mu) \quad , \quad t\epsilon[t_o+\epsilon,t_f] \tag{37}$$

for some to fixed $\epsilon > 0$.

Proof: From (23), we have

$$\int_{t_o}^{t} y(t)dt = \int_{t_o}^{t} [(c_o(t) + O(\mu))p(t) + (c_2(t) + O(\mu))q(t) + w_2(t)]dt$$

$$= \int_{t_o}^{t} [c_o(t)p(t) + c_2(t)q(t) + w_2(t)]dt + O(\mu) \tag{38}$$

since p(t) and q(t) are bounded as $\mu \to o$. Using (26), (27), and lemma 1, (38) becomes

$$\int_{t_o}^{t} y(t)dt = \int_{t_o}^{t} [c_o(t)\eta(t) + D_o(t)u(t) + w_2(t)]dt + O(\mu) \quad .$$

Theorem 1:

The solution of the stochastic linear quadratic problem for the singularly perturbed system defined in (7)-(11), can be replaced to an $O(\mu)$ approximation in J, by the solution of the stochastic linear quadratic problem for the reduced problem defined in (11)-(13).

Proof: From lemma 2, we know that, to an $O(\mu)$ error, the minimization of J reduces to the minimization of J_S with respect to (24). Assuming that $y_1(t)$ is available for measurement, (24), (36) and (34) define a stochastic linear quadratic problem for the slow variable $\eta(t)$, whose solution is given in [8]. Since in that solution $y_1(t)$ is used as an input to a slow a filter, from lemma 3, $y_1(t)$ can be replaced by y(t) giving an $O(\mu)$ error in the estimate $\hat{\eta}(t)$. An $O(\mu)$ error in $\hat{\eta}(t)$ gives an $O(\mu)$ error in J. When $y_1(t)$ is replaced by y(t), the solution of this problem is identical to the solution of the reduced problem defined in (11)-(13). This can be seen by comparing (11), (12) and (13) to (24), (36) and (34) respectively with x_r replaced by η.

Thus from theorem 1 and [8], we obtain the near optimal linear output feedback control

$$u = -(R + D_o'D_o)^{-1}(D_o'C_o + B_o'N)\hat{x}_r \quad , \quad t\epsilon[t_o,t_f] \tag{39}$$

where N(t) is the solution of the matrix Riccati equation

$$\dot{N} = -N(A_o - B_o(R + D_o'D_o)^{-1}D_o'C_o) - (A_o - B_o(R + D_o'D_o)^{-1}D_o'C_o)'N$$

$$+ NB_o(R + D_o'D_o)^{-1}B_o'N - (C_o'C_o - C_o'D_o(R + D_o'D_o)^{-1}D_o'C_o) \tag{40}$$

with terminal condition $N(t_f) = T_1$, and $\hat{x}_r$ is the optimal observer given by

$$\dot{\hat{x}}_r = A_o(t)\hat{x}_r + B_o(t)u(t) + K(t)[y(t) - c_o(t)\hat{x}_r(t) - D_o(t)u(t)] \text{ , } t\epsilon[t_o,t_f] \tag{41}$$

with initial condition $\hat{x}_r(t_o) = E[x(t_o)] = \overline{x}_o$, where

$$K(t) = [\Lambda(t)C_o'(t) + G_1(t)V_{12}(t)] V_2^{-1}(t) \quad , \quad t\epsilon[t_o,t_f] \tag{42}$$

and $\Lambda(t)$ is the solution of the matrix Riccati equation

$$\dot{\Lambda} = (A_o - G_1V_{12}V_2^{-1}C_o)\Lambda + \Lambda(A_o - G_1V_{12}V_2^{-1}C_o)' - \Lambda C_o'V_2^{-1}C_o\Lambda$$

$$+ G_1V_1G_1' - G_1V_{12}V_2^{-1}V_{12}'G_1' \tag{43}$$

with initial condition $\Lambda(t_o) = P_{x_o} = E\{(x-\overline{x}_o)(x-\overline{x}_o)'\}$.

IV. REALIZATION OF THE CONTROL ALGORITHM

A realization of the near optimal control algorithm is given in terms of the matrices of the original system defined in equations (1), (2), (3), (4), and (5).

Let $N = \begin{pmatrix} N_1 & N_2 \\ N_2' & N_3 \end{pmatrix}$, $\Lambda = \begin{pmatrix} \Lambda_1 & \Lambda_2 \\ \Lambda_2' & \Lambda_3 \end{pmatrix}$

$$\tilde{A}_o = \tilde{A}_{11} - \tilde{A}_{12}\tilde{A}_{22}^{-1}\tilde{A}_{21} \quad , \quad \tilde{B}_o = \tilde{B}_1 - \tilde{A}_{12}\tilde{A}_{22}^{-1}\tilde{B}_2 \quad , \quad \tilde{G}_o = \tilde{G}_1 - \tilde{A}_{12}\tilde{A}_{22}^{-1}\tilde{G}_2$$

$$\tilde{C}_o = \tilde{C}_1 - \tilde{C}_2\tilde{A}_{22}^{-1}\tilde{A}_{21} \quad , \quad \tilde{D}_o = -\tilde{C}_2'\tilde{A}_{22}^{-1}\tilde{B}_2 \quad , \quad \tilde{E}_o = -\tilde{C}_2\tilde{A}_{22}^{-1}\tilde{G}_2$$

$$\psi_1 = \Lambda_1\tilde{C}_o' + \Lambda_2\tilde{E}_o' \quad , \quad \psi_2 = \Lambda_2'\tilde{C}_o' + \Lambda_3\tilde{E}_o' \text{ .}$$

Then the near optimal linear control is given by

$$u = (R + \tilde{D}_o'\tilde{D}_o)^{-1} [(\tilde{D}_o'\tilde{C}_o + \tilde{B}_o'N_1)\hat{x}_1 + (\tilde{D}_o'\tilde{E}_o + \tilde{B}_o'N_2)\hat{v}_1] \tag{44}$$

where $\hat{x}_1$ and $\hat{v}_1$ are given by the filters

$$\dot{\hat{x}}_1 = \tilde{A}_o\hat{x}_1 + \tilde{G}_o\hat{v}_1 + \tilde{B}_o u + \psi_1(y - \tilde{c}_o\hat{x}_1 - \tilde{E}_o\hat{v}_1 - \tilde{D}_o u) \quad , \hat{x}_1(t_o) = \bar{x}_{10} \tag{45}$$

$$\dot{\hat{v}}_1 = \tilde{H}\hat{v}_1 + (\psi_2 + V_{12})(y - \tilde{c}_o\hat{x}_1 - \tilde{E}_o\hat{v}_1 - \tilde{D}_o u) \quad , \hat{v}_1(t_o) = \bar{v}_{10} \tag{46}$$

A block diagram of this representation is given in (Fig. 1).

V. DISCUSSION

In the resulting control algorithm, the output measurement y(t) is used as an input to a slow filter only. No filtering of fast variables is done. This permitted us to model the observation noise as white noise. As an illustration, in determining the required speed of response of the measuring instruments only slow variables are of importance. With respect to such low pass instruments, it is justified to model the observation noise as white noise. The assumption of colored input noise and white observation noise is thus consistent with the separation of time scales in singularly perturbed systems.

The result obtained in section III can be considered as a generalization of a similar result obtained in [3] for the deterministic time-invarient case when $t_f = \infty$.

VI. ACKNOWLEDGEMENT

The author wishes to express his gratitude to Professor P. V. Kokotović for his suggestions, guidance and fruitful discussion in the course of this work.

REFERENCES

1) Chang, K. W., "Singular perturbation of a general boundary value problem," Siam J. Math. Anal., Vol. 3, No. 3, August 1972.

2) Chow, J., "Multistage design of singularly perturbed regulators," Proc. 13th Annual Allerton Conference on Circuit and System Theory, 1975.

3) Chow, J., "Separation of time scales in linear time-invariant systems," M.S. Thesis, U. of I., 1975.

4) Haddad, A. H., "Linear filtering of singularly perturbed systems," IEEE Trans. Autom. Control, Vol. AC-21, No. 4, pp. 515-519, August 1976.

5) Haddad, A. H. and Kokotović, P. V., "Stochastic control of linear singularly perturbed systems," This proceeding.

6) Kokotović, P. V. and Sannuti, P., "Singular perturbation method for reducing the model order in optimal control design," IEEE Trans. Autom. Control, Vol. AC-13, No. 4, pp. 377-384, August 1968.

7) Kokotović, P. V. and Yackel, R. A., "Singular perturbation of linear regulators: Basic theorems," IEEE Trans. on Automatic Control, Vol. AC-17, No. 1, pp. 29-37, February 1972.

8) Kwakernaak, H. K. and Sivan, R., "Linear optimal control system," New York, John Wiley, 1972.

9) Sannuti, P., "Asymptotic series solution of singularly perturbed optimal control problems," Automatica, Vol. 10, pp. 183-194, 1974.

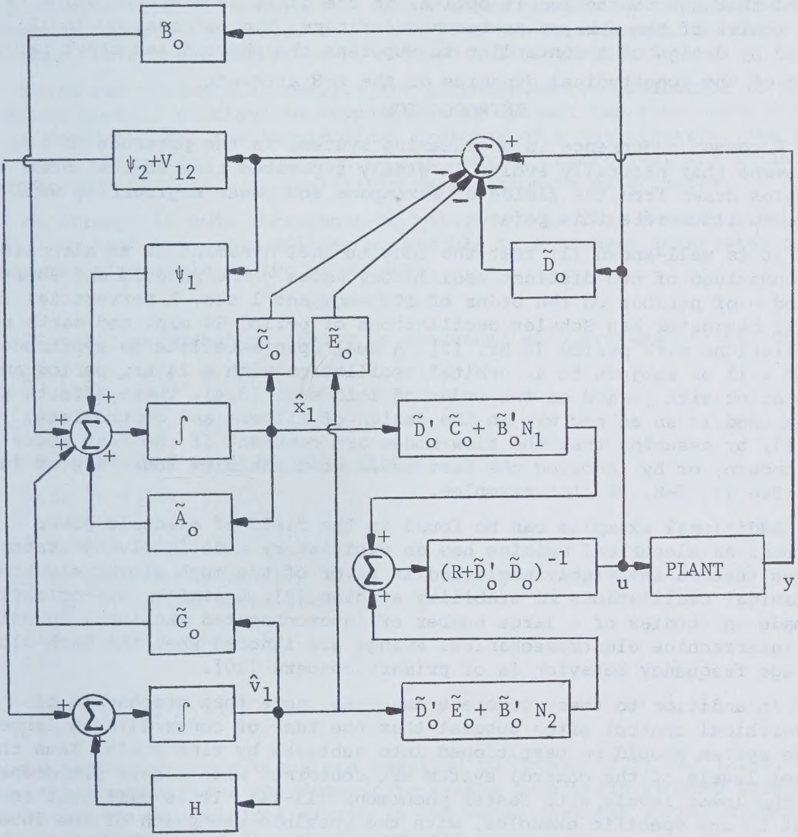

$\tilde{B}_o$
ψ_2+V_{12}
Σ
+
−
−
ψ_1
$\tilde{D}_o$
$\tilde{C}_o$
$\tilde{E}_o$
$\hat{x}_1$
+
+
+
+
Σ
$\int$
$\tilde{D}'_o\ \tilde{C}_o + \tilde{B}'_o N_1$
$\tilde{A}_o$
+
+
Σ
$(R+\tilde{D}'_o\ \tilde{D}_o)^{-1}$
u
PLANT
y
$\tilde{G}_o$
$\hat{v}_1$
+
Σ
$\int$
$\tilde{D}'_o\ \tilde{E}_o + \tilde{B}'_o\ N_2$
H

Fig. (1)

Near Optimal Linear Output Control Algorithm

LINEAR REGULATOR DESIGN FOR STOCHASTIC SYSTEMS BY A MULTIPLE TIME SCALES METHOD*

DEMOSTHENIS TENEKETZIS
Electronic Systems Laboratory
Massachusetts Institute of Technology

NILS R. SANDELL Jr.
Electronic Systems Laboratory
Massachusetts Institute of Technology

ABSTRACT

This paper develops a hierarchically-structured, suboptimal controller for a linear stochastic system composed of fast snd slow subsystems. It is proved that the controller is optimal in the limit as the separation of time scales of the subsystems becomes infinite. The methodology is illustrated by design of a controller to suppress the phugoid and short period modes of the longitudinal dynamics of the F-8 aircraft.

INTRODUCTION

A common occurrence in engineering systems is the presence of phenomena that naturally evolve in widely separated time scales. Some examples drawn from the fields of aerospace and power engineering will serve to illustrate this point.

It is well-known [1] that the longitudinal dynamics of an aircraft are comprised of two distinct oscillatory modes - the phugoid and short period - of periods on the order of 100 sec. and 1 sec. A terrestrial inertial navigator has Schuler oscillations of period 84 min. and earth rate oscillations with period 24 hr. [2]. A dual spin satellite in synchronous orbit will be subject to an orbital oscillation with a 24 hr. period and a nutation with period on the order of 1-10 sec. [3,4]. These effects are often used in an ad hoc way in the design of filters and controllers, usually by assuming that the slow modes are constant if the fast modes are of concern, or by ignoring the fast modes when the slow modes are of interest. See [3, 5-8, 28] for examples.

Additional examples can be found in the field of electric power systems. An electrical machine has an oscillatory mode involving stator fluxes that is invariably neglected in favor of the much slower electromechanical oscillations in stability studies [9]. A similar approximation is made in studies of a large number of interconnected machines, in which the intermachine electromechanical swings are ignored when the much slower average frequency behavior is of primary concern [10].

In addition to these concrete examples, note that proponents of hierarchical control often suggest that the task of controlling a large scale system should be partitioned into subtasks by time scale. Thus the higher levels of the control system are concerned with slower phenomena, and the lower levels with faster phenomena [11-14]. It is difficult to point to any specific examples, with the possible exception of the interaction between automatic generation control and economic dispatch on electric power systems [15].

The multiple time scale phenomena alluded to above are conveniently

---

*This research was conducted at the M.I.T. Electronic Systems Laboratory with support extended by NASA under Grant NGL-22-009-124, and by ERDA under grant ERDA-E(49-18)-2087.

modelled via perturbation theory [16]. There are a number of possible approaches, but we will adopt the framework of singular perturbation theory. This theory has been applied to a variety of control problems by a number of authors [17-21, 24-27] but the previous work most relevant to this paper is that of Kokotovic et. al. [18-20] and Haddad [21]. This paper is in the spirit of [18-21], and moreover requires several of the detailed results of these papers concerning singular perturbations of Ricatti equations.

Specifically, the paper begins by analysis of singular perturbations for linear stochastic systems with two time scales. An approximating system is obtained with the property that the mean-square error between the states of the actual and approximating systems approaches zero as the separation of time scales becomes infinite. The usefulness of this result is demonstrated by application to the stochastic optimal linear regulator problem. With the machinery properly set up, it is straightforward to identify the asymptotically optimal controller, using known results on singular perturbations of Riccati equations. The controller has an interesting hierarchical structure, with the implication of reduced on line computations.

These new theoretical results are illustrated by application to an important control problem. An asymptotically optimal two time scale controller is developed for the longitudinal dynamics of a jet aircraft. The two time scale controller is compared to the optimal controller, and it is demonstrated that there is negligible degradation in performance.

An attempt is made throughout to relegate technical details (see [23]), so that the paper will be accessible to engineers interested only in the two time scale design procedure.

MAIN RESULTS

Singular perturbation theory is concerned with systems of the form

$$\dot{x}(t;\ \varepsilon) = f(x,\ y,\ \varepsilon) \tag{2.1}$$

$$\varepsilon\dot{y}(t;\ \varepsilon) = g(x,\ y,\ \varepsilon) \tag{2.2}$$

and the corresponding degenerate system

$$\dot{x}(t;\ 0) = f(x,\ y,\ 0) \tag{2.3}$$

$$0 = g(x,\ y,\ 0) \tag{2.4}$$

The basic question is whether (2.3) - (2.4) is an approximation to (2.1) - (2.2) in the sense that

$$\lim_{\varepsilon\to 0}\ x(t;\ \varepsilon) = x(t;\ 0) \tag{2.5}$$

$$\lim_{\varepsilon\to 0}\ y(t;\ \varepsilon) = y(t;\ 0) \tag{2.6}$$

Various technical assumptions are required to obtain (2.5) and (2.6), but under these assumptions the degenerate system is a valid reduced order approximation to the original system in the sense that for ε sufficiently small the solutions of the two systems are close.

In the stochastic case, the situation is more complex. Consider the linear system

$$\frac{d}{dt}\begin{bmatrix} x_1(t;\ \varepsilon) \\ \varepsilon x_2(t;\ \varepsilon) \end{bmatrix} = \begin{bmatrix} A_{11}(\varepsilon) & A_{12}(\varepsilon) \\ A_{21}(\varepsilon) & A_{22}(\varepsilon) \end{bmatrix}\begin{bmatrix} x_1(t;\ \varepsilon) \\ x_2(t;\ \varepsilon) \end{bmatrix} + \begin{bmatrix} L_1(\varepsilon) \\ L_2(\varepsilon) \end{bmatrix} \xi(t) \tag{2.7}$$

where

$$E\begin{bmatrix} x_1(0;\ \varepsilon) \\ x_2(0;\ \varepsilon) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{2.8}$$

$$E\begin{bmatrix} x_1(0;\ \varepsilon)x_1^T(0;\ \varepsilon) & x_1(0;\ \varepsilon)x_2^T(0;\ \varepsilon) \\ x_2(0;\ \varepsilon)x_1^T(0;\ \varepsilon) & x_2(0;\ \varepsilon)x_2^T(0;\ \varepsilon) \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{2.9}$$

$$E\{\xi(t)\} = 0 \tag{2.10}$$

$$E\{\xi(t)\xi^T(s)\} = \Xi\delta(t-s) \tag{2.11}$$

$x_1(0)$, $x_2(0)$ are independent of $\xi_1(t)$, $\xi_2(t)$, and all matrices are continuous in ε at $\varepsilon = 0$. Moreover, $A_{22}(0)$ is stable.

An approximation to (2.7) is desired that is valid for small ε and is simpler than (2.7). Note that setting $\varepsilon = 0$ in (2.7) is inadequate; since

$$x_2(t;\ 0) = -A_{22}^{-1}(0)A_{21}(0)x_1(t;\ 0) - A_{22}^{-1}(0)L_2(0)\xi(t) \tag{2.12}$$

has a white noise component and therefore has infinite variance. Consequently,

$$E\left\{\Big(x_2(t;\ \varepsilon) - x_2(t;\ 0)\Big)^T\Big(x_2(t;\ \varepsilon) - x_2(t\ ;\ 0)\Big)\right\} = +\infty \tag{2.13}$$

so that $x_2(t;\ 0)$ is not an approximation to $x_2(t;\ \varepsilon)$ (in the least squares sense).

Instead, define the stochastic degenerate system associated with (2.7) to be the system

$$\dot{x}_{1d}(t;\ \varepsilon) = A_{11d}(\varepsilon)x_{1d}(t;\ \varepsilon) + L_{1d}(\varepsilon)\xi(t),\quad x_{1d}(0;\ \varepsilon) = x_1(0) \tag{2.14}$$

$$\dot{x}_{2d}(t;\ \varepsilon) = A_{21d}(\varepsilon)x_{1d}(t;\ \varepsilon) + A_{22d}(\varepsilon)x_{2d}(t;\ \varepsilon) + L_{2d}(\varepsilon)\xi(t),$$
$$x_{2d}(0;\ \varepsilon) = x_2(0) \tag{2.15}$$

where

$$A_{11d}(\varepsilon) = A_{11}(\varepsilon) - A_{12}(\varepsilon)A_{22}^{-1}(\varepsilon)A_{21}(\varepsilon) \tag{2.16}$$

$$L_{1d}(\varepsilon) = L_1(\varepsilon) - A_{12}(\varepsilon)A_{22}^{-1}(\varepsilon)L_2(\varepsilon) \tag{2.17}$$

$$A_{21d}(\varepsilon) = A_{21}(\varepsilon) \tag{2.18}$$

$$L_{2d}(\varepsilon) = L_2(\varepsilon) \tag{2.19}$$

Notice that the stochastic degenerate system is of the same order as the original system, unlike the situation for deterministic singular perturbations.

Theorem 1

Consider the linear stochastic system (2.7) - (2.11) and a corresponding stochastic degenerate system (2.14) - (2.15). Assume that all matrices in the two systems are continuous in ε at $\varepsilon = 0$, and that $A_{22}(0)$ and $A_{11}(0) - A_{12}(0)A_{22}^{-1}(0)A_{21}(0)$ are stable[1] (i.e., have eigenvalues in the open left half complex plane). Then the stochastic degenerate system is an approximation to the original system in the sense that

$$\lim_{\varepsilon \to 0} E\{(x_1(t;\ \varepsilon) - x_{1d}(t;\ \varepsilon))(x_1(t;\ \varepsilon) - x_{1d}(t;\ \varepsilon))^T\} = 0 \tag{2.22}$$

$$\lim_{\varepsilon \to 0} E\{(x_2(t;\ \varepsilon) - x_{2d}(t;\ \varepsilon))(x_2(t;\ \varepsilon) - x_{2d}(t;\ \varepsilon))^T\} = 0 \tag{2.23}$$

uniformly for $0 \leq t < \infty$.

Proof

The proof of the theorem is quite involved. Differential equations for

$$\Sigma_{11}(t;\varepsilon) \overset{\Delta}{=} E\{\tilde{x}_1(t;\varepsilon)\tilde{x}_1^T(t;\varepsilon)\} \overset{\Delta}{=} E\{(x_1(t;\varepsilon)-x_{1d}(t;\varepsilon))(x_1(t;\varepsilon)-x_{1d}(t;\varepsilon))^T\} \tag{2.24}$$

$$\Sigma_{22}(t;\varepsilon) \overset{\Delta}{=} E\{\tilde{x}_2(t;\varepsilon)\tilde{x}_2^T(t;\varepsilon)\} \overset{\Delta}{=} E\{(x_2(t;\varepsilon)-x_{2d}(t;\varepsilon))(x_2(t;\varepsilon)-x_{2d}(t;\varepsilon))^T\} \tag{2.25}$$

are obtained, and the limits are evaluated by (non-stochastic) singular perturbation theory to establish (2.22) and (2.23). See [23] for details.

Remarks

1. Note that the stochastic degenerate system is the same order as the original system, so that (2.23) is valid for $t = 0$.

2. Clearly, the assumption $A_{11}(0) - A_{12}(0)A_{22}^{-1}(0)A_{21}(0)$ stable is only necessary to insure uniform convergence in (2.22), (2.23) on the infinite interval. Without this assumption, a theorem of Tihonov [16] can be invoked which insures uniform convergence in (2.22), (2.23) for sets of the form $[0,T]$.

3. Proof is easily generalized to cover uniformly asymptotically stable time-varying systems at the expense of some additional notation.

Consider now the system

$$\dot{x}(t;\ \varepsilon) = A(\varepsilon)x(t;\ \varepsilon) + B(\varepsilon)u(t) + L(\varepsilon)\xi(t) \tag{2.26}$$

with observations

$$y(t;\ \varepsilon) = Cx(t;\ \varepsilon) + \theta(t) \tag{2.27}$$

and cost

$$J = \lim_{T \to \infty} \frac{1}{T} [E\{\int_0^T x^TQx + u^TRu\ dt\}] \tag{2.28}$$

---

[1] Note that this assumption implies that there exists an $\varepsilon_0 > 0$ such that the system matrix in (2.7) is stable for all $0 \leq \varepsilon < \varepsilon_0$ [22].

where

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad A = \begin{bmatrix} A_{11} & A_{12} \\ \frac{A_{21}}{\varepsilon} & \frac{A_{22}}{\varepsilon} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ \frac{B_2}{\varepsilon} \end{bmatrix}$$

$$L = \begin{bmatrix} L_1 \\ \frac{L_2}{\varepsilon} \end{bmatrix}, \quad C = \begin{bmatrix} C_1 & C_2 \end{bmatrix}, \quad Q = \begin{bmatrix} Q_{11} & Q_{21} \\ Q_{21} & Q_{22} \end{bmatrix} \geq 0, \quad R > 0$$

$$E\{\xi(t)\} = 0, \qquad E\{\theta(t)\} = 0$$

$$E\{\xi(t)\xi^T(s)\} = \Xi\delta(t-s), \quad E\{\theta(t)\theta^T(s)\} = \Theta\delta(t-s)$$

$$\Xi > 0, \qquad \Theta > 0$$

and $\xi(t)$, $\theta(t)$ are independent and Gaussian. The assumptions

$$[A(\varepsilon), B(\varepsilon)], \quad [A(\varepsilon), L(\varepsilon)] \quad \text{controllable} \tag{2.29}$$

$$[A(\varepsilon), C], \quad [A(\varepsilon), \sqrt{Q}] \quad \text{observable} \tag{2.30}$$

are made, $0 < \varepsilon \leq \varepsilon_0$.

As is well known, the optimal control law is

$$u(t;\ \varepsilon) = -\ G(\varepsilon)\hat{x}(t;\ \varepsilon) \tag{2.31}$$

where

$$G(\varepsilon) = R^{-1}B^T K(\varepsilon) \tag{2.32}$$

and $K(\varepsilon)$ satisfies

$$0 = -\ K(\varepsilon)A(\varepsilon) - A^T(\varepsilon)K(\varepsilon) - Q(\varepsilon) + K(\varepsilon)B(\varepsilon)R^{-1}B^T(\varepsilon)K(\varepsilon) \tag{2.33}$$

The estimate satisfies the equation

$$\dot{\hat{x}}(t;\ \varepsilon) = A(\varepsilon)\hat{x}(t;\ \varepsilon) + H(\varepsilon)(y(t;\ \varepsilon) - C\hat{x}(t;\ \varepsilon)) + B(\varepsilon)u(t;\ \varepsilon) \tag{2.34}$$

where

$$H(\varepsilon) = \Sigma(\varepsilon)C'\Theta^{-1} \tag{2.35}$$

and $\Sigma(\varepsilon)$ satisfies

$$0 = \Sigma(\varepsilon)A^T(\varepsilon) + A(\varepsilon)\Sigma(\varepsilon) + L(\varepsilon)\Xi L^T(\varepsilon) - \Sigma(\varepsilon)C^T R^{-1} C\Sigma(\varepsilon) \tag{2.36}$$

At this point, we are ready to apply Theorem 1 to approximate the controller (2.31), (2.34) by a two time scale controller. Because of Theorem 1, any system that has a stochastic degenerate system in common with the optimal closed-loop system will be asympotically optimal. The optimal closed-loop system can be written

$$\begin{bmatrix} \dot{x}_1(t;\varepsilon) \\ \dot{\hat{x}}_1(t;\varepsilon) \\ \varepsilon\dot{x}_2(t;\varepsilon) \\ \varepsilon\dot{\hat{x}}_2(t;\varepsilon) \end{bmatrix} =$$

$$\begin{bmatrix} A_{11} & -B_1G_1(\varepsilon) & A_{12} & -B_1G_2(\varepsilon) \\ H_1(\varepsilon)C_1 & A_{11}-B_1G_1(\varepsilon)-H_1(\varepsilon)C_1 & H_1(\varepsilon)C_2 & A_{12}-B_1G_2(\varepsilon)-H_1(\varepsilon)C_2 \\ A_{21} & -B_2G_1(\varepsilon) & A_{22} & -B_2G_2(\varepsilon) \\ \varepsilon H_2(\varepsilon)C_1 & A_{21}-B_2G_1(\varepsilon)-\varepsilon H_2(\varepsilon)C_1 & \varepsilon H_2(\varepsilon)C_2 & A_{22}-B_2G_2(\varepsilon)-\varepsilon H_2(\varepsilon) \end{bmatrix}$$

$$\begin{bmatrix} x_1(t;\varepsilon) \\ \hat{x}_1(t;\varepsilon) \\ \hat{x}_2(t;\varepsilon) \\ \hat{x}_2(t;\varepsilon) \end{bmatrix} + \begin{bmatrix} L_1 & 0 \\ 0 & H_1(\varepsilon) \\ L_2 & 0 \\ 0 & \varepsilon H_2(\varepsilon) \end{bmatrix} \begin{bmatrix} \xi(t) \\ \theta(t) \end{bmatrix} \tag{2.37}$$

Note that the stochastic degenerate system can be obtained by eliminating x_2, $\hat{x}_2$ from the equations for x_1, $\hat{x}_1$ using the algebraic relations that result when ε is set equal to zero in the left hand side of (2.37). Of course, the resulting system cannot be implemented since the value of ε in the x_1 and x_2 equations is not a design parameter.

An implementable system that has a stochastic degenerate system in common with (2.37) is obtained as follows. Assume that $(A_{22}-B_2G_2(\varepsilon)-\tilde{H}_2(\varepsilon)C_2)^{-1}$ exists. Set $\varepsilon = 0$ in the left hand side of only the $\hat{x}_2$ equations of (2.37) to obtain:

$$\hat{x}_2 = -(A_{22} - B_2G_2(\varepsilon) - \tilde{H}_2(\varepsilon)C_2)^{-1}[(A_{21}-B_2G_1(\varepsilon) - \tilde{H}_2(\varepsilon)C_1)\hat{x}_1 + \\ + \tilde{H}_2(\varepsilon)C_1x_1 + \tilde{H}_2(\varepsilon)C_2x_2 + \tilde{H}_2(\varepsilon)\theta] \tag{2.38}$$

Substitute into the $\hat{x}_1$ equation to obtain

$$\dot{\hat{x}}_{1d}(t;\ \varepsilon) = A_{11D}\hat{x}_{1d}(t;\ \varepsilon) + H_{1D}y(t;\ \varepsilon) \tag{2.39}$$

where

$$A_{11D}(\varepsilon) = A_{11}-B_1G_1(\varepsilon) - H_1(\varepsilon)C_1 - (A_{12}-B_1G_2(\varepsilon)-H_1(\varepsilon)C_2) \times \\ (A_{22}-B_2G_2(\varepsilon)-\tilde{H}_2(\varepsilon)C_2)^{-1}(A_{21} - B_2G_1(\varepsilon)- \tilde{H}_2(\varepsilon)C_1) \tag{2.40}$$

$$H_{1D}(\varepsilon) = H_1(\varepsilon)-(A_{12}-B_1G_2(\varepsilon)-H_1(\varepsilon)C_2)(A_{22}-B_2G_2(\varepsilon)-\tilde{H}_2(\varepsilon)C_2)^{-1}\tilde{H}_2(\varepsilon) \tag{2.41}$$

$$\tilde{H}_2(\varepsilon) = \varepsilon H_2(\varepsilon) \tag{2.42}$$

Based on this analysis, the following suboptimal closed-loop system is obtained.

$$\begin{bmatrix} \dot{x}_{1D}(t;\varepsilon) \\ \dot{\hat{x}}_{1D}(t;\varepsilon) \\ \varepsilon\dot{x}_{2D}(t;\varepsilon) \\ \varepsilon\dot{\hat{x}}_{2D}(t;\varepsilon) \end{bmatrix} =$$

$$\begin{bmatrix} A_{11} & -B_1G_1(\varepsilon) & A_{12} & -B_1G_2(\varepsilon) \\ H_{1D}(\varepsilon)C_1 & A_{11D}(\varepsilon) & H_{1D}(\varepsilon)C_2 & 0 \\ A_{21} & -B_2G_1(\varepsilon) & A_{22} & -B_2G_2(\varepsilon) \\ \tilde{H}_2(\varepsilon)C_1 & A_{21}-B_2G_1(\varepsilon)-\tilde{H}_2(\varepsilon)C_1 & \tilde{H}_2(\varepsilon)C_2 & A_{22}-B_2G_2(\varepsilon)-\tilde{H}_2(\varepsilon)C_2 \end{bmatrix}$$

$$\begin{bmatrix} x_{1D}(t;\varepsilon) \\ \hat{x}_{1D}(t;\varepsilon) \\ x_{2D}(t;\varepsilon) \\ \hat{x}_{2D}(t;\varepsilon) \end{bmatrix} + \begin{bmatrix} L_1 & 0 \\ 0 & H_{1D}(\varepsilon) \\ L_2 & 0 \\ 0 & \tilde{H}_2(\varepsilon) \end{bmatrix} \begin{bmatrix} \xi(t) \\ \theta(t) \end{bmatrix} \tag{2.43}$$

Clearly, (2.43) is not the stochastic degenerate system of (2.37), but has been constructed to have a stochastic degenerate system in common with (2.37)

Theorem 2.

Consider the LQG problem defined by (2.26) - (2.28). Assume that the controllability and observability assumptions (2.29), (2.30) hold and that $(A_{22} - B_2G_2(\varepsilon)-H_2(\varepsilon)C_2)^{-1}$ exists, $0< \varepsilon \leq \varepsilon_0$. Then the suboptimal closed-loop system (2.43) is asymptotically optimum in the following sense:

$$\lim_{\varepsilon\to 0} E\{(x_1(t;\varepsilon) - x_{1D}(t;\varepsilon))(x_1(t;\varepsilon)-x_{1D}(t;\varepsilon))^T\} = 0$$

$$\lim_{\varepsilon\to 0} E\{(\hat{x}_1(t;\varepsilon)-\hat{x}_{1D}(t;\varepsilon))(\hat{x}_1(t;\varepsilon)-\hat{x}_{1D}(t;\varepsilon))^T\} = 0$$

$$\lim_{\varepsilon\to 0} E\{(x_2(t;\varepsilon)-x_{2D}(t;\varepsilon))(x_2(t;\varepsilon)-x_{2D}(t;\varepsilon))^T\} = 0$$

$$\lim_{\varepsilon\to 0} E\{(\hat{x}_2(t;\varepsilon)-\hat{x}_{2D}(t;\varepsilon))(\hat{x}_2(t;\varepsilon)-\hat{x}_{2D}(t;\varepsilon))^T\} = 0$$

Proof

As noted above, (2.37) and (2.43) have a common stochastic degenerate system which governs their behavior as $\varepsilon \to 0$. Therefore, all that is required is the verification of the hypotheses of Theorem 1.

The required continuity properties follow from [18, 20]. Note that under the stated controllability and observability assumptions the degenerate and boundary layer system controllability and observability conditions required are satisfied [24]. The required stability properties are established as a consequence of the stability of the closed loop LQG design and a previously quoted result [22]. Full details can be found in [23].

Remarks

1. Note that the above results are not the most general possible, since the time-varying and finite horizon case could probably also be solved. However, the above results are of the greatest practical interst.

2. Extension to more than two time scales in straightforward [22].

3.Investigation of the rather ad hoc assumption $(A_{22}-B_2(\varepsilon)G_2(\varepsilon)-H_2(\varepsilon)C_2)^{-1}$ exists for $0 < \varepsilon \leq \varepsilon_0$ would be of theoretical interest since verification is difficult. As a practical matter, this issue is less important, since invertibility for the value of ε of interest is easily checked. The performance of the two time scale controller can then be directly assessed.

4. The proposed suboptimal controller is illustrated in Figure 1. Notice that there is a unidirectional interface between the slow and fast filters. Thus there is opportunity for considerable reduction in on-line computational effort since the two sets of filter equations can be numerically integrated in different time scales (i.e., with different step sizes).

5. The above development assumed, for simplicity, that the optimal gain matrices $H(\varepsilon)$, $G(\varepsilon)$ were computed. In fact, by a more elaborate analysis, it is possible to show that only asymptotic approximations to $H(\varepsilon)$, $G(\varepsilon)$ near $\varepsilon = 0$ need to be computed [22]. It is possible to compute these approximations by the methods extensively studied in [18-21], thus obtaining a potential reduction in off-line as well as on-line computation.

6. Notice that it is the two time scale nature of the closed loop system that is required. In the above analysis, as $\varepsilon \to 0$ the closed loop system automatically has fast and slow modes. In a physical system ε has a fixed, non-zero value. Consequently, it is possible that the open loop dynamics can have fast and slow modes, but in the closed loop these slow modes are eliminated.

TWO TIME SCALE CONTROL OF AIRCRAFT LONGITUDINAL DYNAMICS

The particular problem to be addressed in this section is the design of a feedback control system for the longitudinal dynamics of an F-8 aircraft. Specifically the controller must produce elevator commands to keep the aircraft in steady level flight in the face of wind disturbances. For simplicity, the wind disturbance is modelled as white.

The equations of motion of an airplane are a set of coupled nonlinear equations in the longitudinal and lateral state variables. If the equations are linearized about nominal state and control variables, then the resulting linear equations are found to approximately decouple into separate sets for the longitudinal and lateral dynamics. See [1] for an excellent discussion of the modelling issues.

The aircraft's longitudinal variables are

$$x = \begin{bmatrix} V \\ \gamma \\ \alpha \\ q \end{bmatrix}, \quad u = \delta_e$$

where

V = horizontal velocity deviation in ft/sec
γ = flight path angle in radians,
α = angle of attack in radians,
q = pitch rate in rad/sec.
δ_e = is the elevator deflection in radians

The interpretation of these variables is given in Figure 2. Table 1 gives the system matrices. It is assumed that velocity and pitch rate measurements, both corrupted by wideband noise, are available.

Figures 3-4 show the system response to an initial pitch $\theta(0) = 1^0$, $V(0) = 100$ ft/sec in the absence of the wind disturbance. The two time scale behavior is well illustrated here. Table 2 gives the system's eigenvalues and eigenvectors. Note that the variables V, γ dominate the slower phugoid mode, and the variables α, q dominate the faster short period mode. The physical nature of these oscillations is beautifully described on pages 320-328 of [1].

From the above discussion and inspection of the system A matrix, the equations for V and γ are the logical candidates for the slow dynamics and the equations for α and q are suggested as the fast dynamics. For determination of ε, the following procedure is suggested. Note that the state equations can be written

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ \varepsilon x_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ \varepsilon A_{21} & \varepsilon A_{22} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ \varepsilon B_2 \end{bmatrix} u + \begin{bmatrix} L_1 \\ \varepsilon L_2 \end{bmatrix}$$

where $x_1^T = [V\ \gamma]$, $x_2^T = [\alpha\ q]$. For $\varepsilon = .01$, the matrix

$$\varepsilon A_{22} \begin{bmatrix} -.012 & .0100 \\ -.0901 & .0069 \end{bmatrix}$$

has eigenvalues comparable to those of A_{11}. Determination of a value of ε is actually not required for design of the two time scale controller, but is useful for judging the appropriateness of the approximation $\varepsilon = 0$.

Before proceeding with the regulator design, a few remarks are in order. First, the aircraft longitudinal variables more often include the pitch θ instead of the flight path angle γ. It was only after considerable difficulty that the above formulation, in which the choice of fast and slow variables is clear, was hit upon. Second, the extensive literature on singular perturbation theory contains almost no discussion of the choice of fast and slow variables, or the choice of ε. But determination of these quantities is the first problem one has to face up to in applications of the theory.

A LQG controller was designed for the Q and R matrices Q = diag [.01,0,3260,3260], R = [3260] using standard routines [29]. Design goals were to (i) achieve a damping ratio $\zeta > .707$ for both modes (ii) reduce the state variables response to the wind disturbance. Closed-loop eigenvalues and RMS state variable and estimation error standard deviations are given in Table 3. As noted in remark 6 of Theorem 2, it is critical that the open loop separation of modes be present also in the closed loop. From Table 2, the optimal design has this property. The closed-loop eigenvalues of the two time scale controller and corresponding RMS state variable and estimation error standard deviations are also listed in Table 3. Note the generally good correspondance between the optimal and suboptimal designs.

SUMMARY AND CONCLUSIONS

This paper has considered the reduction in on-line computational effort for an LQG design with fast and slow closed loop modes. Together with the results of [18-21], this paper demonstrates that the singular perturbation approach to the LQG problem offers the potential of near optimal performance with reduction in both on-and off-line computation.

The design procedure of this paper has been applied to control of the longitudinal dynamics of a jet aircraft. A two time scale design was obtained with performance extremely close to that of the optimal design. Note that recent proposals for adaptive flight control systems require multiple Kalman filters running in parallel [30, 31], so that reduction of on-line computation is of definite interest.

Several directions for future research are evident. First, procedures for systematically picking the fast and slow variables of a system with fast and slow modes, as well as for determining explicitly ε, would be highly useful in applications. Second, note that the design of Figure 1 has an interesting hierarchical structure. In fact, as pointed out in [14], if a system is composed of a number of fast subsystems with a slow interconnecting equation,a decentralized, hierarchical design is naturally obtained. Therefore, the results of this paper are of potential interest in hierarchical systems theory. Finally, note that singular perturbation theory is only one approach to the multiple time scale phenomenon. The results of Ramnath [32], for example, provide another approach that could be exploited in control theory.

REFERENCES

1. B. Etkin, Dynamics of Atmospheric Flight, Wiley, N.Y., 1972.
2. C. Broxmeyer, Inertial Navigation Systems, McGraw-Hill, New York, 1964.
3. C.R. Johnson, "TACSAT I Nutation Dynamics", in N.E. Feldman and C.M. Kelley, Eds., Communication Satellites for the 70's: Technology, M.I.T. Press, Cambridge, Massachusetts, 1971.
4. Y.C. Tao, Satellite Attitude Prediction by Multiple Time Scales Method, Ph.D. Thesis, M.I.T., Feb. 1976.
5. W.H. Lee, K.P. Dunn, M. Athans, "A Study of the Continuous Time LQG Problem for a Reduced Model of the F-8 Aircraft Longitudunal Dynamics", ESL Interim Report 6, NASA Grant NSG-1018, M.I.T., March 1975.
6. W.F. O'Halloran, Jr. and R. Warren, "Design of a Reduced State Suboptimal Filter for Self-Calibration of a Terrestrial Inertial Navigation System", AIAA Guidance and Control Conference, Stanford University, 1972.
7. A.E. Bryson, Jr., "Rapid In-Flight Estimation of IMU Platform Misalignments Using External Position and Velocity Data", Unpublished memo, Stanford University, January 1973.
8. W.L. Black, B. Howland, E.A. Brablick, "An Electromagnetic Attitude Control System for a Synchronous Satellite", AIAA J. Spacecraft, Vol. 6, No. 7, July 1969.
9. D.W. Olive, "Digital Simulation of Synchronous Machine Transients", IEEE Trans. P.A.S., August, 1968.
10. M.L. Chan, R.D. Dunlop, F.C. Schweppe, "Dynamic Equivalents for Average System Frequency Behavior Following Major Disturbances", IEEE Trans. P.A.S., Vol. PAS-91, No. 4, July/August 1972.
11. M.D. Mesarovic, D. Macko, and Y. Takahara, Theory of Hierarchical, Multilevel Systems, Academic Press, New York, 1970.
12. C.Y. Chong, On the Decentralized Control of Large Scale Systems, Ph.D. Thesis, M.I.T., 1972.
13. F.C. Schweppe and S.K. Mitter, "Hierarchcal Systems Theory and Electric Power Systems", in E. Handschin, Ed., Real-time Control of Electric Power Systems, Elsevier, 1972.
14. N.R. Sandell, Jr., P. Varaiya, and M. Athans,"Survey of Decentralized Control Methods for Large Scale Systems", Proc. Engineering Foundation Conference on Systems Engineering for Power: Status and Prospects, Henniker, New Hampshire, August 1975.
15. J. Zaborszky, A.K. Subramarian, K.M. Lu, "Control Interfaces in

Generation Allocation", Proc. Engineering Foundation Conference on Systems Engineering for Power: Status and Prospects, Henniker, New Hampshire, August 1975.

16. W.W. Wasow, Asymptotic Expansions for Ordinary Differential Equations, Interscience, New York, 1965.

17. Singular Perturbations: Order Reduction in Control System Design, American Society of Mechanical Engineers, New York, 1972.

18. P.P. Kokotovic and R.A. Yackel, "Singular Perturbation of Linear Regulators" Basic Theorems", IEEE Trans. A.C., Vol. AC-17, No. 1, February 1972.

19. A.H. Haddad and P.V. Kokotovic, "On a Singular Perturbation Problem in Linear Filtering Theory", Proc. 5th Annual Princeton Conference on Information Sciences and Systems, Princeton, N.J., March 1971.

20. A.H. Haddad and P.V. Kokotovic, "On Singular Perturbations in Linear Filtering and Smoothing," Proc. 5th Symposium on Nonlinear Estimation Theory and Its Applications, San Diego, CA. September 1974.

21. A.H. Haddad, "Linear Filtering of Singularly Perturbed Systems", IEEE Trans. A.C., Vol. AC-21, No. 4, August 1976.

22. C.A. Desoer and M.J. Shensa, "Networks with Very Small and Very Large Parasitics: Natural Frequencies and Stability", Electronic Research Laboratory, University of California, Berkeley, Memo No. ERL-4276, April 1970.

23. D. Teneketzis, "Perturbation Methods in Decentralized Stochastic Control," S.M. Thesis, M.I.T., Cambridge, Massachusetts 1976.

24. P.V. Kokotovic and A.H. Haddad, "Controllability and Time Optimal Control of Systems with Slow and Fast Modes", IEEE Trans. A.C., Vol. AC-20, No. 1, February 1975.

25. T.N. Edelbaum, "Singular Perturbations and Minimum Fuel Space Trajectories", Eighth Annual Allerton Conference on Circuit and System Theory, Monticello, Illinois, Oct. 1970.

26. P.V. Kokotovic, R.E. O'Malley Jr., and P. Sannuti, "Singular Perturbations and Order Reduction in Control Theory - An Overview", Sixth World IFAC Conference, Cambridge, Massachusetts, Sept. 1975.

27. A.J. Calise, "Singular Perturbation Methods for Variational Problems in Aircraft Flight", IEEE Trans. A.C., Vol. AC-21, No. 3, June 1976.

28. M.F. Barrett and G. Stein, "Performance Predictions for Gyro-Based Satellite Attitude Control Systems at Geosynchronous Altitude", AIAA Guidance and Control Conference, San Diego, Calif., August 1976.

29. N.R. Sandell Jr. and M. Athans, Modern Control Theory:Computer Manual, M.I.T. Center for Advanced Engineering Study, Cambridge, Mass., 1974.

30. M. Athans et. al., "The Stochastic Control of the F-8C Aircraft Using the Multiple Model Adaptive Control (MMAC) Method", to appear in IEEE Trans. A.C.

31. G. Stein, G. Hartman, R. Hendrick, "Adaptive laws for F-8 Flight Test", to appear in IEEE Trans. A.C.

32. R.V. Ramnath and G. Sundri, "A Generalized Multiple Time Scales Approach to a Class of Linear Differential Equations", J. Math. Anal. Ap. Vol. 28, No. 2, November 1969.

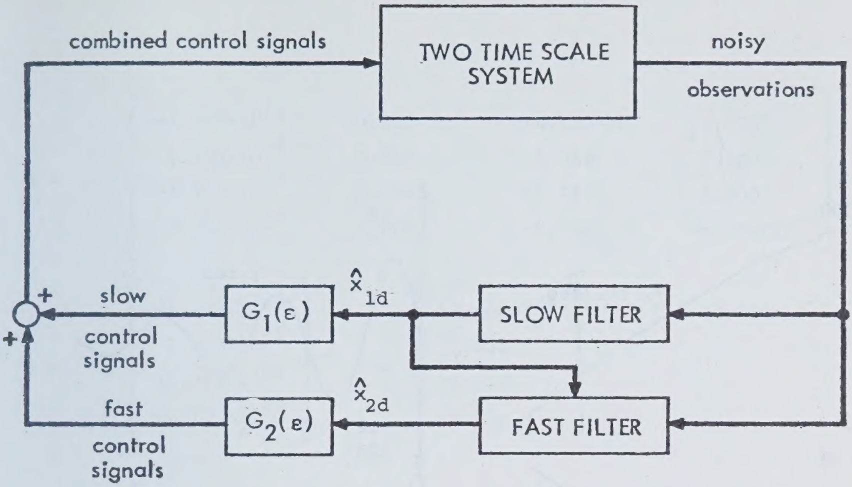

combined control signals
TWO TIME SCALE SYSTEM
noisy observations
$\hat{x}_{1d}$
SLOW FILTER
$G_1(\varepsilon)$
slow control signals
+
+
$\hat{x}_{2d}$
FAST FILTER
$G_2(\varepsilon)$
fast control signals

Fig. 1 Asymptotically Optimal Two Time Scale Controller

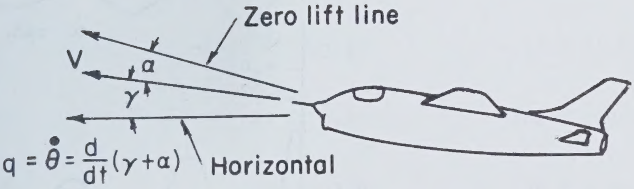

Zero lift line
V
α
γ
$q = \dot{\theta} = \frac{d}{dt}(\gamma+\alpha)$
Horizontal

Figure 2 Aircraft Longitudinal Variables

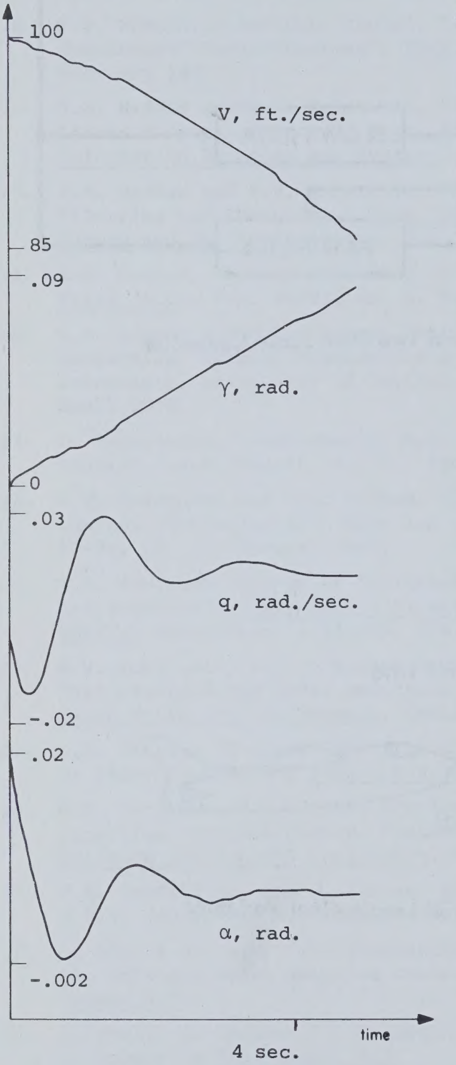


Figure 3.Response in Fast Time Scale

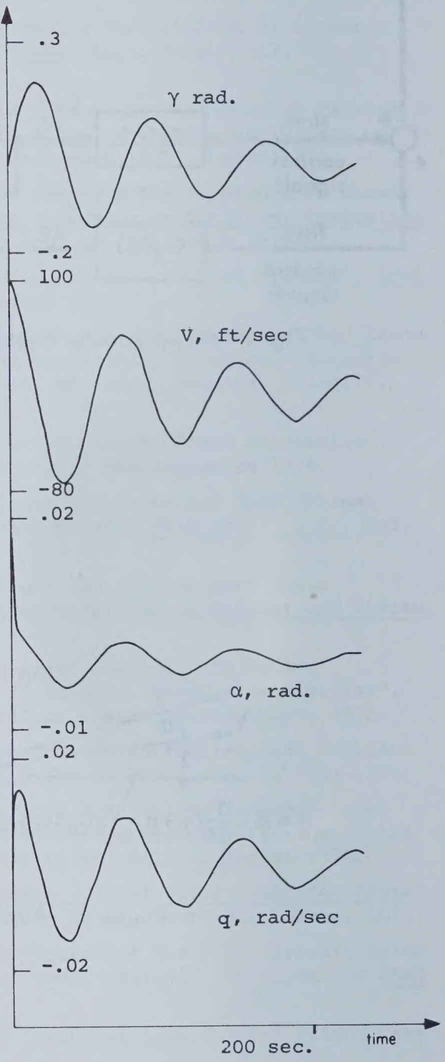


Figure 4.Response in Slow Time Scale

$$A = \begin{bmatrix} -1.357\times10^{-2} & -3.220\times10^{1} & -4.630\times10^{1} & 0.000 \\ 1.200\times10^{-4} & 0.000 & 1.214 & 0.000 \\ -1.212\times10^{-4} & 0.000 & -1.214 & 1.000 \\ 5.700\times10^{-4} & 0.000 & -9.010 & -6.696\times10^{-1} \end{bmatrix}$$

$$B = \begin{bmatrix} -4.330\times10^{-1} \\ 1.394\times10^{-1} \\ -1.394\times10^{-1} \\ -1.577\times10^{-1} \end{bmatrix} \qquad L = \begin{bmatrix} -4.630\times10^{1} \\ 1.214 \\ -1.214 \\ -9.010 \end{bmatrix} \qquad \Xi = [3.150\times10^{-4}]$$

$$C = \begin{bmatrix} 0.0 & 0.0 & 0.0 & 1.0 \\ 1.0 & 0.0 & 0.0 & 0.0 \end{bmatrix} \qquad \Theta = \begin{bmatrix} 6.859\times10^{-4} & 0.000 \\ 0.000 & 4.000\times10^{1} \end{bmatrix}$$

Table 1. System Matrices

eigenvalue		$-.94 \pm j.2.98$	$-.0075 \pm j.076$
eigenvector components	v	$1.3 \pm j.4.7$	$-4.3\times10^{2} \mp j1.1\times10^{2}$
	γ	$-5\times10^{-4} \mp j.13$	$-.15 \pm j1.0$
	α	$.33 \pm j.10$	$-.02 \mp j.005$
	q	$-.21 \pm j1.0$	$-.08 \mp j.02$

Table 2. Open Loop Eigenvalues
and Eigenvectors.

Eigenvalues

open loop	-.94 ± j2.98	-.075 ± j0.76		
optimal	-2.9 ± 2.0	-.20 ± j.20	-3.8, -2.6	-.09 ± j.10
two time scale	-3.1 ± j2.3	-.18 ± j.16	-3.8, -2.6	-.1 ± j.11

RMS Deviations

	$\sqrt{\overline{V^2}}$	$\sqrt{\overline{\gamma^2}}$	$\sqrt{\overline{\alpha^2}}$	$\sqrt{\overline{q^2}}$	$\sqrt{\overline{(V-\hat{V})^2}}$	$\sqrt{\overline{(\gamma-\hat{\gamma})^2}}$	$\sqrt{\overline{(\alpha-\hat{\alpha})^2}}$	$\sqrt{\overline{(q-\hat{q})^2}}$
open loop	8.8	.0024	.0031	.0082				
optimal	3.3	.0018	.0026	.0065	2.6	.0013	.0021	.0056
two time scale	3.4	.0018	.0026	.0065	2.6	.0015	.0021	.0056

Table 3. Performance Comparison

A THEOREM CONCERNING PROPERTIES OF AN ARMS-RACE MODEL

I. W. SANDBERG
Mathematical Studies Department
Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

In an earlier paper some analytical results and associated interpretations are given concerning the dependence on parameters of the equilibrium solutions of a class of n-country nonlinear arms-race models. The hypothesis that the dynamic equation of evolution has an equilibrium point that possesses a certain stability property plays an important role in that study. Here a proof is given of the proposition that under certain reasonable and interesting conditions, the solution of the usual version of the equation of evolution in fact always approaches such a point. Some related additional material is presented as well. The results are useful also in connection with models of several other types of social processes including, for example, Homans-type group interaction processes.

I. INTRODUCTION

In Ref. 1 some analytical results and associated interpretations are given concerning the dependence on parameters of the equilibrium solutions of a class of n-country nonlinear nonsubmissive arms-race models of the Richardson type.[2] It is proved, for example, that if certain hypotheses are satisfied, then, roughly speaking, hate is at least as costly in an economic sense (as far as percentage changes in expenditures for armaments are concerned) to those who hate as it is to those who are hated.[†]

The differential-equation version of the model considered in Ref. 1 is of the form

$$\frac{dx}{dt} = f(x), \quad x(0) = x_0, \quad t \geq 0 \tag{1}$$

in which x is a real vector of dimension n and f has certain special properties.

In our discussion concerning (1), we adopt the convention that if v is any n-vector, then v_i denotes the i^{th} component of v. In (1), for each i, x_i denotes the cost of armaments[#] of Country i. Due to the assumption of nonsubmissiveness, f has the off-diagonal monotonicity property that for each i, $f_i(u) \geq f_i(v)$ for any two armaments vectors u and v that lie in the domain of definition of f and satisfy $u_i = v_i$ and $u_j \geq v_j$ for $j \neq i$.

The hypothesis that f has an equilibrium point which possesses a certain one-sided stability property plays a key role in Ref. 1. The main purpose of this note is to give a proof of the proposition that under certain reasonable conditions, the solution of (1) always approaches such a point. Some related additional material is also presented.

[†] See Ref. 1 for the details. An earlier special (i.e., local) result along similar lines is contained in Ref. 3.

[#] In some studies the x_i denote levels (or stocks) of armaments, rather than the associated costs.

Although applications other than the arms-race model application are not explicitly discussed, the results are useful (and have analogous interpretations) in connection with models of several other types of social processes including, for instance, the various Homans-type group interaction processes described in Ref. 3.

II. DESCRIPTION AND APPLICATIONS OF THE THEOREM PROVED IN THE APPENDIX

Let r and s be two real n-vectors such that $r_i \leq s_i$ for all i, and such that for each i all conceivable expenditures for armaments of interest to us in connection with Country i lie in the interval $[r_i, s_i]$. Let D denote the set of all n-vectors whose i^{th} component lies in $[r_i, s_i]$ for each i.

We assume that D is the domain of f, and that on D, f is continuous and has the "off-diagonal monotonicity" property described in the preceding section. It is also assumed that f satisfies the Lipschitz condition described in the Appendix.

We suppose that there are two n-vectors u_0 and w_0 which lie in D and have the property that $(u_0)_i \leq (w_0)_i$ and $f_i(u_0) \geq 0 \geq f_i(w_0)$ for all i. It is reasonable to assume that two such vectors exist on the grounds that if the armaments level of every country is sufficiently small or sufficiently large, respectively, there will be no tendency for any country to decrease or to increase its armaments level. In fact, it would be reasonable to require that

$$r_i < s_i \quad \text{and} \quad f_i(r) > 0 \geq f_i(s) \quad \text{for each } i, \tag{2}$$

but this stronger hypothesis is not necessary for the purpose of proving the theorem we wish to describe.[†] However, in so far as the particular application considered in this section is concerned, we shall assume that (2) holds.

According to Part (i) of the theorem proved in the Appendix, (1) possesses a unique solution (which must of course be D-valued and hence bounded) for any starting point x_0 that lies in D.[#]

[†] The theorem has some worth-while applications of a related but less global type than described here. As a very simple example, if we set $u_0 = r$ and $w_0 = s$, it asserts the existence, under certain conditions, of an equilibrium point in D that is stable relative to D in the sense of Ref. 1 even if the n-dimensional interval D is instead very much smaller than the set of all conceivable armaments-expenditure vectors of interest to us.

[#] For a proof, choose u_0 and w_0 of the theorem to be r and s, respectively. It appears that the result described above, which in particular directs attention to the boundedness of the solution under certain conditions, has not appeared in the arms-race literature, even for the linear case considered by Richardson. Simple examples show that the result would be false if we had omitted the condition that f has the off-diagonal monotonicity property.

According to Part (ii) of the theorem, for each $x_0 \in D$ either $x(t)$ of (1) converges as $t \to \infty$ to a point x^* in D such that $f_i(x^*) = 0$ for all i (i.e., to an equilibrium point in D), or D contains at least two equilibrium points. This proposition is analogous to a result proved in Ref. 4 concerning a model which in several ways is considerably more complex. It has significant applications along the lines discussed in Sections 4.12 and 4.13 of Ref. 4.[†]

At this point it is convenient to introduce the following notation. With u and v any two elements of D, the inequality $v \geq u$ or $v > u$, respectively, means that $v_i \geq u_i$ for all i or $v_i > u_i$ for all i, and if $v \geq u$, we denote by $[u,v]$ the set $\{z \in D: u \leq z \leq v\}$. The zero element of the set of all n-vectors is denoted by θ.

In order to present our result concerning the one-sided stability condition, which involves the initial condition x_0 of (1), let r' be any element of D such that $r' > r$ and

$$x_0 < s \quad \text{and} \quad f(x_0) > \theta \geq f(s) \tag{3}$$

whenever $r' \geq x_0 \geq r$. Since (2) holds and f is continuous, it is clear that a suitable r' exists. Let X_0 denote $[r,r']$, which is of course simply a set of initial conditions of (1) such that x_0 is sufficiently close to r in the sense that (3) holds.

We shall use Part (iii) of the theorem to show that if $x_0 \in X_0$, then $x(t)$ approaches an equilibrium point u^* in D as $t \to \infty$, and u^* has the property that for any $v \in D$ such that $v \leq u^*$ and $v \neq u^*$, we have $f_k(v) > 0$ for some k. This is our main result. Roughly speaking, it shows that if the arms race starts from a point at which every country has some tendency to increase its level of armaments, and certain other conditions are satisfied, then the arms race will converge to an equilibrium point that satisfies a "stability condition" that appears as an important hypothesis in Ref. 1.

In order to show the validity of the result stated in the preceding paragraph we use the fact that according to Part (iii) of the theorem, the assertion is true if in the preceding paragraph "D" is replaced with $[x_0,s]$ (i.e., with $[u_0,w_0]$ with $u_0 = x_0$ and $w_0 = s$). Moreover if $y \in ([r,s]-[x_0,s])$, then $r_k \leq y_k < r_k'$ for some k, and hence, using $f_k(z) > 0$ for all $z \in X_0$ as well as the off-diagonal monotonicity of f, we have $f_k(y) > 0$. Therefore, as claimed, u^* satisfies the condition that for any $v \in D$ such that $v \leq u^*$ and $v \neq u^*$, we have $f_k(v) > 0$ for some k.

[†] Of course the proposition implies that if there is exactly one equilibrium point in D, then $x(t)$ converges to that point as $t \to \infty$ for each $x_0 \in D$. Simple examples show that the proposition would be false even if we had defined f on the set R^n of all real n-vectors, and had let D be the set of all initial-condition vectors, if we had omitted, for example, the hypothesis that $f_i(r) > 0$ for all i or the hypothesis that f has the off-diagonal monotonicity property on R^n.

APPENDIX: THE THEOREM

Notation and Definitions

Let n be an arbitrary positive integer, and let R^n denote the set of all real n-vectors. For each $v \in R^n$, let v_i denote[†] the i^{th} component of v for each i, and let $\|v\|$ denote $\sum_{i=1}^{n} |v_i|$. For $v \in R^n$ and $u \in R^n$, the inequality $v \geq u$ means that $v_i \geq u_i$ for all i. The zero vector of R^n is denoted by θ.

Let r and s be elements of R^n such that $s \geq r$, and let D denote the n-dimensional interval $\{z \in R^n: r \leq z \leq s\}$.

Let f be a continuous mapping of D into R^n with the property that for each i, $f_i(u) \geq f_i(v)$ whenever $u \in D$, $v \in D$, $u \geq v$, and $u_i = v_i$. Suppose also that, with ρ a positive constant, f satisfies the Lipschitz condition $\|f(u)-f(v)\| \leq \rho \|u-v\|$ for all $u \in D$ and all $v \in D$.

Let w_0 and u_0 be elements of D such that $w_0 \geq u_0$ and $f(u_0) \geq \theta \geq f(w_0)$. Let $[u_0,w_0]$ denote $\{z \in D: u_0 \leq z \leq w_0\}$.

By an equilibrium point in $[u_0,w_0]$ is meant a $v \in [u_0,w_0]$ such that $f(v) = \theta$.

Consider the differential equation (1). As usual, by a solution of (1) we mean a function that takes values in D, is defined and differentiable with respect to t on $[0,\infty)$, and satisfies (1) for all $t \geq 0$.

Theorem. Let $x_0 \in D$ satisfy $w_0 \geq x_0 \geq u_0$. Then

(i) There is exactly one solution x(t) of (1).

(ii) Either there is an equilibrium point $x^* \in [u_0,w_0]$ such that $x(t) \to x^*$ as $t \to \infty$, or there are at least two equilibrium points in $[u_0,w_0]$.

(iii) If $x_0 = u_0$, then $[u_0,w_0]$ contains an equilibrium point u^* such that $x(t) \to u^*$ as $t \to \infty$, and u^* has the property that for any $v \in [u_0,w_0]$ such that $v \leq u^*$ and $v \neq u^*$, we have $f_k(v) > 0$ for some index k. Similarly, if $x_0 = w_0$, then $[u_0,w_0]$ contains an equilibrium point w^* such that $x(t) \to w^*$ as $t \to \infty$, and w^* has the property that for any $v \in [u_0,w_0]$ such that $v \geq w^*$ and $v \neq w^*$, we have $f_k(v) < 0$ for some k. In addition, $u^* \leq w^*$ and any equilibrium point p in $[u_0,w_0]$ satisfies $u^* \leq p \leq w^*$.

Proof: Let g be the mapping of R^n into D defined by the condition that for each $v \in R^n$ and for each i, $g_i(v) = v_i$ if $r_i \leq v_i \leq s_i$, $g_i(v) = r_i$ if $v_i < r_i$, and $g_i(v) = s_i$ if $v_i > s_i$. We see that the composite mapping fg is continuous on R^n and satisfies the Lipschitz condition $\|f[g(u)]-f[g(v)]\| \leq \rho\|u-v\|$ for $u \in R^n$ and $v \in R^n$. Thus, by a well-known result (see, for example, Ref. 5) there is a unique solution y of

$$\frac{dy}{dt} = f[g(y)], \quad y(0) = x_0, \quad t \geq 0. \tag{4}$$

[†] Some of the preliminary material of Section II is repeated here for the convenience of the reader.

Therefore to prove Part (i) of the theorem, it suffices to show that $w_0 \geq x_0 \geq u_0$ implies that $y(t) \in D$ for all $t \geq 0$. We shall use the following four lemmas.

Lemma 1. If $u \in R^n$, $v \in R^n$, and $u \geq v$, and if δ is any positive constant such that $\delta < \rho^{-1}$, then

$$u + \delta f[g(u)] \geq v + \delta f[g(v)].$$

Proof of Lemma 1:

Since it is clear that for each i, $|f_i[g(u)]-f_i[g(v)]| \leq \rho|u_i-v_i|$ for the special case in which $u \geq v$ and $u_j = v_j$ for $j \neq i$, Lemma 1 is in fact proved in Appendix I of Ref. 4.

The following lemma is proved by Kantorovich.[6]

Lemma 2. Let G be a mapping of R^n into itself such that $G(u) \geq G(v)$ whenever $u \geq v$, and suppose that $u_0 \leq w_0$, $u_0 \leq G(u_0)$, and $w_0 \geq G(w_0)$. Then the sequences $u_1, u_2, \ldots$ and $w_1, w_2, \ldots$ defined by

$$u_{k+1} = G(u_k), \; w_{k+1} = G(w_k), \; k = 0,1,\ldots$$

satisfy $u_k \leq u_{k+1} \leq w_0$ and $w_k \geq w_{k+1} \geq u_0$ for all k.[†]

Lemma 3. Let t_1 and t_2 be positive numbers such that $t_2 > t_1$. Then, with u(t) and w(t), respectively, the solution y(t) of (4) corresponding to $x_0 = u_0$ and $x_0 = w_0$, we have $w(t_1) \geq w(t_2) \geq u_0$ and $u(t_1) \leq u(t_2) \leq w_0$.

Proof of Lemma 3:

Suppose that w(t) does not satisfy $w(t_1) \geq w(t_2) \geq u_0$, which means there is an index k such that at least one of the inequalities $w_k(t_1) < w_k(t_2)$ and $w_k(t_2) < (u_0)_k$ holds.

Let $w_{(p)}(t)$ denote the continuous piecewise-linear approximation to w(t), on the interval $[0,t_2]$, obtained by interpolating in the usual way between the successive points generated by Euler's formula

$$z_{k+1} = z_k + hf[g(z_k)], \; z_0 = w_0, \; k = 0,1,\ldots$$

with step size $h = t_2(2^p)^{-1}$. It is well known that $w_{(p)}(t)$ converges uniformly to w(t) on $[0,t_2]$ as $p \to \infty$.[#] Moreover, using Lemmas 1 and 2, we have $w_{(p)}(t_1) \geq w_{(p)}(t_2) \geq u_0$ whenever $t_2(2^p)^{-1} < \rho^{-1}$, which contradicts both $w_k(t_1) < w_k(t_2)$ and $w_k(t_2) < (u_0)_k$. Therefore, $w(t_1) \geq w(t_2) \geq u_0$. A similar argument shows that $u(t_1) \leq u(t_2) \leq w_0$. This completes the proof of Lemma 3.

[†] Kantorovich[6] proved also that when G is continuous on R^n, the limits u_* and w_*, respectively, of $\{u_k\}$ and $\{w_k\}$ satisfy $u_* = G(u_*)$ and $w_* = G(w_*)$, and any solution q of $u_0 \leq q \leq w_0$ and $q = G(q)$ satisfies $u_* \leq q \leq w_*$. We do not use this fact in our proof, but direct attention to it in order to show the relation of our theorem to earlier work.

[#] See, for example, Ref. 7 and recall that fg is continuous and satisfies a Lipschitz condition on R^n.

Lemma 4. Let u(t), x(t), and w(t), respectively, denote the solution y(t) of (4) corresponding to $x_0 = u_0$, $x_0 = x_0$, and $x_0 = w_0$. Then $u(t) \leq x(t) \leq w(t)$ for all $t \geq 0$.

Proof of Lemma 4:

The lemma follows from a result given in Ref. 8 and the fact that (4), with off-diagonally-monotone f, possesses a unique solution for each initial condition in D.

Since Lemmas 3 and 4 show that $y(t) \in D$ for all $t \geq 0$ (in fact that $y(t) \in [u_0,w_0]$ for all t), we have proved Part (i).

By Lemmas 3 and 4, there exist points u* and w* such that $u(t) \to u^*$ as $t \to \infty$, $w(t) \to w^*$ as $t \to \infty$, and $u_0 \leq u^* \leq w^* \leq w_0$. Since f is continuous on D, $f(u^*) = f(w^*) = \theta$. By Lemma 4, if there is no equilibrium point $x^* \in [u_0,w_0]$ such that $x(t) \to x^*$ as $t \to \infty$, then $u^* \neq w^*$, which proves Part (ii).

Suppose (for the sake of obtaining a contradiction) that $f(v) \leq \theta$ for some $v \in [u_0,w_0]$ such that $v_j < (u^*)_j$ for at least one value of j. Then since $u_0 \leq v$ and $f(u_0) \geq \theta \geq f(v)$, it follows from the material presented above that the solution of (4) with $x_0 = u_0$ converges to an equilibrium point p such that $u_0 \leq p \leq v$, which is a contradiction. Therefore u* has the property that any equilibrium point p in $[u_0,w_0]$ satisfies $u^* \leq p$, and for each $v \in [u_0,w_0]$ such that $v \leq u^*$ and $v \neq u^*$, we have $f_k(v) > 0$ for some k. Of course, a similar argument shows that w* has the properties stated in the theorem.

This completes the proof of the theorem.

REFERENCES

1. Sandberg, I. W., "Some Qualitative Properties of Nonlinear Richardson-Type Arms Race Models," to appear in Mathematical Systems in International Relations Research, J. Gillespie and D. Zinnes (eds.).

2. Richardson, L. F., Arms and Insecurity, Chicago: Quadrangle Books, 1960.

3. Sandberg, I. W., "On the Mathematical Theory of Interaction in Social Groups," IEEE Trans. on Systems, Man, and Cybernetics, SMC-4, 5(1974), 432-445.

4. Sandberg, I. W., "On the Mathematical Theory of Social Processes Characterized by Weak Reciprocity," to appear.

5. Nemytskii, V., and Stepanov, V., Qualitative Theory of Differential Equations, Princeton: Princeton University Press, 1960, 9-13.

6. Kantorovich, L., "The Method of Successive Approximations for Functional Equations," Acta Math. 71(1939), 63-97.

7. Henrici, P., Discrete Variable Methods in Ordinary Differential Equations, New York: John Wiley, 1968, 112-116.

8. Lakshmikantham, V., and Leela, S., Differential and Integral Inequalities, Vol. I, New York: Academic Press, 1969, 28-29.

A PIECEWISE-LINEAR INPUT-OUTPUT MODEL

MING-JEH CHIEN
Wayne State University
Detroit, Michigan 48202

ABSTRACT

This paper presents a piecewise-linear version of the well-known Input-Output model of multisectored economy of Leontief. This model does not require that the functions describing the input-output technology are continuously differentiable and/or concave (or convex). The main theorem shows that the model is workable in the sense that for any nonnegative vector of final demands there is a unique nonnegative vector of production levels of sectors of the economy. Furthermore, higher demand requires higher production level.

I. INTRODUCTION

Input-Output analysis of a multisectored economy, developed originally by Leontief, is an economist's approach to describe quantitatively the inter-sectored (interindustry) relationships in an economic structure [1,2]. A basic phenomenon of the input-output model is that the inputs to a particular sector (industry) are outputs from a number of other sectors (industries). In other words, the industrial outputs are used not only to satisfy consumer demands, but also as intermediate products to satisfy the needs of other sectors (industries). Let the quantity of the output of sector i needed to produce x_j units of the output of sector j is $a_{ij}x_j$. Then the model is characterized by the equation

$$x - Ax = y \tag{1}$$

where x is the vector of production levels of sectors and y is the vector of final demands. The entries of the square matrix A describe the relationships between sectors. Conceptually, the definition of the matrix is straight-forward. However, one should be cautioned about the implicit assumptions. First, with constant a_{ij}'s, the input-output model is assumed to be linear and time-invariant. Linearity implies constant return to scale. Second, the fixed proportion of inputs implies no substitutability which is not always true in a competitive economy. Moreover, the input-output model implies no joint products, i.e., an output x_j is only produced by sector j.

In this paper the relationships between sectors are assumed to be piecewise linear. More specifically, the equation describing the economy is

$$x - A(\cdot)x = y \tag{2}$$

The (i,j)th entry of the operator $A(\cdot)$, $a_{ij}(\cdot)$ is a continuous piecewise-linear function of x_j [3-9]. This model allows both increasing and decreasing return to scale. It also allows substitutability. For example, let the amounts of x_j and x_k needed for the production of x_i be specified by $a_{ji}(x_i)$ and $a_{ki}(x_i)$ respectively. When the production level of x_i is changed, the needed amounts of x_j and x_k can be changed without being at fixed proportions. In other words, certain amount of x_j might be used to substitute x_k according to the market situation. Nonetheless, this model still assumes no joint products to simplify the analysis.

The proof of the main theorem relies on properties of a special class of matrices, which is referred to as Class K matrices and is defined in Section II.

II. THE MODEL

2.1 Preliminaries

In this paper, a square matrix is a real function on I x I, where I is some index set. Let a set of indices 1, 2, ..., n be denoted by N. If $M \subset N$ and if A is a matrix on N x N, the partial function of A on M x M is denoted by A(M) which is called the principal submatrix of A corresponding to M. Let det A be the determinant of a matrix A. The number det A(M) is called the principal minor of A corresponding to M. The (i,j)th entry of a matrix A is denoted by a_{ij}. Although an n-dimensional column vector can be considered as a n x 1 matrix, we should use x_i to denote the ith (or (i,1)th) entry of a vector x.

Definition 1: Let Z denote the class of real square matrixes whose off-diagonal elements are all nonpositive.
Definition 2: Let P denote the class of real square matrixes whose principal minors are all positive.
Definition 3: Let K denote the class of real square matrixes which are in both Z and P.
Definition 4: A function f(x) is called "affine" if it has the form f(x) = Ax + w, where A and w are constant real matrix and vector respectively.
Definition 5: A continuous function f(x) is said to be piecewise-linear if the domain of the function is divided into a finite number of nonoverlapping closed convex polyhedrons (regions), and the function is affine in every polyhedron (region).
Definition 6: Region R_k is said to be effective at x if $x \in R_k$.

A typical continuous piecewise linear function from R^1 to R^1 is shown in Fig. 1.

Theorem A[10]: Let $A \in K$. If $B \geqq A$ and $B \in Z$ then

$$\text{(i) } B \in K, \text{ (ii) } 0 \leqq B^{-1} \leqq A^{-1} \text{ and (iii) } \det B \geqq \det A > 0$$

Theorem B[7]: Let f(•) from R^n (n-dimensional Euclidean space) into itself be continuous piecewise-linear. Then f(•) is a homeomorphism(*) if at any x, the convex combination(**) of Jacobian matrices of effective regions at x is nonsingular.

This theorem is easily proven by the fact that a continuous piecewise-linear function is a homeomorphism if it is one-to-one and onto [9].

(*) A mapping $f(\cdot): D \subset R^n \to R^n$ is a homeomorphism of D onto f(D) if f(•) is one-to-one on D and f(•) and $f^{-1}(\cdot)$ are continuous on D and f(D) respectively.

(**) The convex combination of matrices $J_1, J_2, \ldots, J_m$, is defined as $\sum_{i=1}^{m} \mu_i J_i$ where $1 \geqq \mu_i \geqq 0$, $i = 1, 2, \ldots, m$ and $\sum_{i=1}^{m} \mu_i = 1$.

2.2 The Piecewise-Linear Input-Output Model

The (i,j)th entry of the operator A(·)of equation (2), $a_{ij}(\cdot)$ is a continuous piecewise-linear function from R^1 into itself, namely,

$$a_{ij}(x_i) = s_{ij}(k_{ij})\ x_j + w_{ij}(k_{ij}),\ k_{ij} = 1, 2, \ldots, K_{ij}. \qquad (3)$$

Where K_{ij} is the number of segments (the segments are consecutively numbered from left to right, i.e., $-\infty$ to $+\infty$). For example, in Fig. 2, $s_{ij}(1) = 0$, w_{ij}, $(1) = 0$, $s_{ij}\ (2) = 2$, $w_{ij}\ (2) = 0$, $s_{ij}\ (3) = 1$ and $w_{ij}\ (3) = 1$. There are three segments and $K_{ij} = 3$.

From $a_{ij}(\cdot)$'s, a continuous piecewise-linear function is completely specified. Example 1: Let

$$a_{11}(x_1) = \begin{cases} s_{11}(1)\ x_1 + w_{11}(1) & \text{for } -\infty \leqq x_1 \leqq 1 \\ s_{11}(2)\ x_i + w_{11}(2) & \text{for } 1 \leqq x_1 \leqq \infty \end{cases}$$

$$a_{21}(x_1) = \begin{cases} s_{21}(1)\ x_1 + w_{21}(1) & \text{for } -\infty \leqq x_1 \leqq 0 \\ s_{21}(2)\ x_1 + w_{21}(2) & \text{for } 0 \leqq x_1 \leqq \infty \end{cases}$$

$$a_{12}(x_2) = \begin{cases} s_{12}(1)\ x_2 + w_{12}(1) & \text{for } -\infty \leqq x_2 \leqq 0 \\ s_{12}(2)\ x_2 + w_{12}(1) & \text{for } 0 \leqq x_2 \leqq \infty \end{cases}$$

$$a_{22}(x_2) = \begin{cases} s_{22}(1)\ x_2 + w_{22}(1) & \text{for } -\infty \leqq x_2 \leqq 0 \\ s_{22}(2)\ x_2 + w_{22}(2) & \text{for } 0 \leqq x_2 \leqq \infty \end{cases}$$

From equation (2) the above functions specify a continuous piecewise-linear function which has 6 regions, namely,

Region I:	$-\infty \leqq x_1 \leqq 0$,	$-\infty \leqq x_2 \leqq 0$
Region II:	$0 \leqq x_1 \leqq 1$,	$-\infty \leqq x_2 \leqq 0$
Region III:	$1 \leqq x_1 \leqq \infty$,	$-\infty \leqq x_2 \leqq 0$
Region IV:	$-\infty \leqq x_1 \leqq 0$,	$0 \leqq x_2 \leqq \infty$
Region V:	$0 \leqq x_1 \leqq 1$,	$0 \leqq x_2 \leqq \infty$
Region VI:	$1 \leqq x_1 \leqq \infty$,	$0 \leqq x_2 \leqq \infty$

The affine function in a region can be easily obtained. For example, the function in Region II is:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} (s_{11}(1)\ x_1 + w_{11}(1)) + (s_{12}(1)\ x_2 + w_{12}(1)) \\ (s_{21}(2)\ x_1 + w_{21}(2)) + (s_{22}(1)\ x_2 + w_{22}(1)) \end{pmatrix}, \text{ or}$$

$$\left[I - \begin{pmatrix} s_{11}(1) & s_{11}(1) \\ s_{21}(2) & s_{22}(1) \end{pmatrix} \right] \begin{bmatrix} x_1 \\ x_x \end{bmatrix} - \begin{pmatrix} w_{11}(1) + w_{12}(1) \\ w_{21}(2) + w_{22}(1) \end{pmatrix}.$$

Since the model is to describe a multisectored economy, we should make the following assumptions.

Assumption 1: For all i, j,

$$a_{ij}(x_j) = 0 \qquad\qquad x_j < 0$$

$$a_{ij}(x_j) \geqq 0 \qquad\qquad x_j \geqq 0$$

This asserts that the amount of input required to produce nonnegative output is always nonnegative. This property is essential from an economic point of view. Since the present model is an extension of the linear model, it is desirable to preserve certain properties that a linear model has. We should, therefore, further assume that more output requires more input in any sector.

Assumption 2: For all i, j,

$$a_{ij}(k_{ij}) \geqq 0 \qquad\qquad k_{ij} = 1, 2, \ldots, K_{ij}$$

Let vector $e = [1, 1, \ldots, 1]^T$

$$\text{matrix } S_q = \begin{bmatrix} s_{11}(k_{11}{}^q) & \cdots & s_{1n}(k_{1n}{}^q) \\ \vdots & & \vdots \\ s_{n1}(k_{n1}{}^q) & \cdots & s_{nn}(k_{nn}{}^q) \end{bmatrix},$$

$$\text{matrix } W_q = \begin{bmatrix} w_{11}(k_{11}{}^q) & \cdots & w_{1n}(k_{1n}{}^q) \\ \vdots & & \vdots \\ w_{n1}(k_{n1}{}^q) & \cdots & w_{nn}(k_{nn}{}^q) \end{bmatrix},$$

$$\text{matrix } Y = \begin{bmatrix} y_1 & 0 & \cdots & 0 \\ 0 & y_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & & y_n \end{bmatrix}$$

Equation (2) can then be written as

$$(I - S_q)\, x = (Y + W_q)e, \quad q = 1, 2, \ldots Q \tag{4}$$

where I is the identity matrix, and Q is the number of regions.

Assumption 3: Let $s_{ij}{}^m = \max_{k_{ij}} \{s_{ij}(k_{ij})\}$.

$$\text{The matrix } B^m = I - \begin{pmatrix} s_{11}{}^m & \cdots & s_{1n}{}^m \\ \vdots & & \vdots \\ s_{n1}{}^m & \cdots & s_{nn}{}^m \end{pmatrix} \text{ belongs to Class P.}$$

From the above assumptions, we can derive the following.

Lemma 1: $B_q = I - S_q$ is a Class K matrix, $q = 1, 2, \ldots, Q$ and $B_q^{-1} \geq 0$

Proof: From Assumptions 2 and 3, $B^m \varepsilon K$, $B_q \varepsilon Z$ and $B_q \geq B^m$

From Theorem A, $B_q \varepsilon K$ and $B_q^{-1} \geq 0$.

Theorem 1: Under Assumptions 1, 2 and 3, the mapping $(I - A(\cdot)): R^n \to R^n$ is a homeomorphism. Furthermore the solution of (4), $x^* \varepsilon R_+^n$ if the demand vector $y \varepsilon R_+^n$, where $R_+^n = \{x \mid x_i \geq 0,$ $i = 1, 2, \ldots, n\}$.

Proof: Let $B_1, \ldots, B_m$ be the Jacobian matrices of the effective regions at x.

Since $B_i \geq B^m$ $i = 1, 2, \ldots, m$, the convex combination of B_1, B_2, ... and B_m,

$$\sum_{i=1}^{m} \mu_i B_i \geq \left(\sum_{i=1}^{m} \mu_i \right) B^m = B^m.$$

Hence $\sum_{i=1}^{m} \mu_i B_i \varepsilon K$ and is nonsingular.

From Theorem B the mapping $(I - A(\cdot))$ is a homeomorphism. We should prove $x^* \varepsilon R_+^n$ by contradiction.

Let $D = \{i \mid x^*_i < 0\}$ and $\overline{D} = \{j \mid x^*_j \geq 0\}$.

Suppose D is not empty. Let $i \varepsilon D$. The ith equation of (2) is

$$x^*_i - \sum_{j=1}^{m} a_{ij} (x^*_j) = y_i.$$

Now

$$x^*_i - \sum_{j=1}^{m} a_{ij}(x^*_j) = x^*_i - \sum_{j \varepsilon D} a_{ij}(x^*_j) - \sum_{j \varepsilon \overline{D}} a_{ij}(x^*_j)$$

$$= x^*_i - \sum_{j \varepsilon \overline{D}} a_{ij}(x^*_j) \leq x^*_i$$

from Assumptions 1 and 2.
Therefore $y_i \leq x^*_i < 0$. This is a contradiction because $y \varepsilon R_+^n$.
This completes the proof.

Corollary 1: Let $\overline{y}$ and y be two demand vectors, and $\overline{x}$ and x be the unique vectors of production levels respectively. If $\overline{y} \geq y$ then $\overline{x} \geq x$.

Proof: Let the inverse mapping of $(I - A(\cdot))$ be $(I - A(\cdot))^{-1}$ which is continuous piecewise-linear from Theorem 1. Let $y(\lambda) = y + \lambda(\overline{y} - y)$, $\lambda \varepsilon [0,1]$.
It is easy to see that $y(\lambda) \geq y$ since $\overline{y} - y \geq 0$. The image of $y(\lambda)$ under $(I - A(\cdot))^{-1}$ is a continuous function of λ and

$(I - A(\cdot))^{-1}(y(\lambda)) = x$ if $\lambda = 0$ and $(I - A(\cdot))^{-1}(y(\lambda)) = \overline{x}$ if $\lambda = 1$. Let $\lambda_1 \varepsilon[0,1]$ be the "largest" number such that $(I - A(\cdot))^{-1}(y(\lambda))$ is in the same region (polyhedron) for $\lambda \varepsilon[0, \lambda_1]$.

Then from equation (4),

$$x(\lambda) = B_q^{-1} (Y(\lambda) + W_q)\, e \quad \text{for } \lambda\varepsilon[0, \lambda_1].$$

Since $x(\lambda_1) - x = B_q^{-1} (y(\lambda_1) - y)$ and $y(\lambda_1) \geqq y$,

$x(\lambda_1) \geqq x$ from Lemma 1.

This procedure can be continued and $\overline{y}$ is reached in a finite number of steps because the number of regions (polyhedrons) is finite. In other words, we can compute $\lambda_1, \lambda_2, \ldots, \lambda_m$ such that $y \leqq y(\lambda_1) \leqq y(\lambda_2) \leqq \ldots \leqq y(\lambda_m) \leqq \overline{y}$ and $x \leqq x(\lambda_1) \leqq x(\lambda_2) \leqq \ldots \leqq x(\lambda_m) \leqq \overline{x}$, as shown in Fig. 3.

Corollary 1 shows that the model has a very desirable property, namely, higher demand would require, higher production level.

A special case of this model is when

$$w_{ij}(k_{ij}) \geqq 0 \quad \text{for all } i \text{ and } j$$

Assumption $\overline{2}$: For all i and j.

$$s_{ij}(k_{ij}) \geqq 0 \text{ and } w_{ij}(k_{ij}) \geqq 0, \quad k_{ij} = 1, 2, \ldots, K_{ij}$$

Lemma 2: Under Assumptions 1 and $\overline{2}$,

$$s_{ij}^{m} = s_{ij}(2) > 0, \quad \text{otherwise}$$

$$s_{ij}^{m} = 0 \text{ and } a_{ij}(x_j) = 0 \quad \text{for all } x_j.$$

Proof: From Assumption 1, it is obvious that $a_{ij}(x_j) = 0$ for all x_j if $a_{ij}(\cdot)$ has only one segment. Let $a_{ij}(\cdot)$ have more than one segment. Since $a_{ij}(\cdot)$ is continuous $w_{ij}(2) = 0$ and $s_{ij}(2) > 0$. If there is $s_{ij}(3)$ then $s_{ij}(3) < s_{ij}(2)$, since $w_{ij}(3) \geqq 0$. Let $\overline{x}_j$ be the intersection of segment 3 and segment 4 (if exists). Let r be the slope of the straight line connecting $\overline{x}_j$ and the origin. It is clear that $s_{ij}(4) \leqq r \leqq s_{ij}(2)$, otherwise $w_{ij}(4)$ will intersect the vertical axis at a negative value. Hence Lemma 2 is shown by induction because the number of segments is finite.

Under Assumptions 1, $\overline{2}$ and 3, the proof of the second half Theorem 1 follows directly from a property of Class K matrices.
For $y \varepsilon R_+^n$, the unique solution

$$x^* = B_q^{-1} (Y + W_q)\, e.$$

From Lemma 1, $B_q^{-1} \geqq 0$, hence,

$$x^* \geqq 0$$

because $Y \geqq 0$ and $W_q \geqq 0$ by assumption.

III. CONCLUSION

The behavior of a continuous piecewise-linear function and algorithms to locate a solution have been extensively developed. The advantages of using piecewise-linear models are as follows:

(i) It does not require differentiability which usually complicates the analysis.

(ii) A piecewise-linear model is compatible with tables of sample data, which are the most convenient way to describe the behavior between sectors (industries).

(iii) The algorithms can locate a solution in a finite number of steps and, at the same time, plot continuously the variation of production level according to the change of demand.

In this paper, a piecewise-linear input-output model is proposed. The model is shown workable in the sense that nonnegative demand will require nonnegative production level. Furthermore, higher demand would result in higher production level. The model is versatile in that it does not assume that the relationship between sectors is concave or convex. We conclude that it is highly feasible to use piecewise-linear model in the analysis of input-output multisectored economy.

REFERENCES

1. W. W. Leontief, "Input Output Economics". New York: Oxford Univ. Press, 1966.

2. W. W. Leontief, "Input-Output in Environmental Modeling", Keynote Address presented at the 1972 IEEE Conf. Decision and Control, New Orleans, La., Dec. 13-15, 1972.

3. M. J. Chien and E. S. Kuh, "Solving Piecewise-Linear Equations for Resistive Networks", International Journal of Circuit Theory and Applications, Vol. 4, pp. 3-24, Jan., 1976.

4. M. J. Chien, "Piecewise-Linear Theory and Computation of Nonlinear Resistive Networks", Ph.D. Dissertation, Department of EECS, U.C. Berkeley, March 1975.

5. I. W. Sandberg, "A Nonlinear Input-Output Model of a Multisectored Economy", Econometrica, Vol. 41, No. 6, (Nov. 1973).

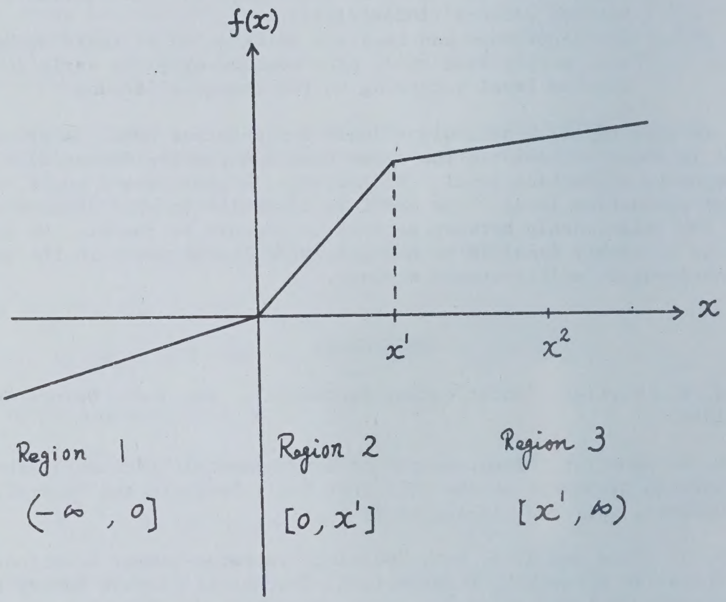
6. I. W. Sandberg, "Some Note on the Optimization of the Gross National Product of Economies with Nonlinear Input-Output Technology", IEEE Trans. on Automatic Control, Oct. 1974.

7. M. J. Chien, "Conditions for a Piecewise-Linear Resistive Network Function to be Homeomorphic", Proceedings of the Eighth Southeastern Symposium on System Theory, pp. 102-107, April 1976.

8. H. Scarf, "The Computation of Economic Equilibria", New Haven and London, Yale University Press, 1973, pp. 37-50 and Chapter 7.

9. T. Fujisawa and E. S. Kuh, "Piecewise-Linear Theory of Nonlinear Networks", SIAM J. Appl. Math., Vol. 22, No. 2, March 1972.

10. M. Fiedler and V. Ptak, "Some Results on Matrices of Class K and Their Application to the Convergence Rate of Interation Procedures", Czech. Math. Journal 16, 260-273 (1966).



Effective regions at x^1: 2 and 3

Effective region at x^2: 3

Fig. 1

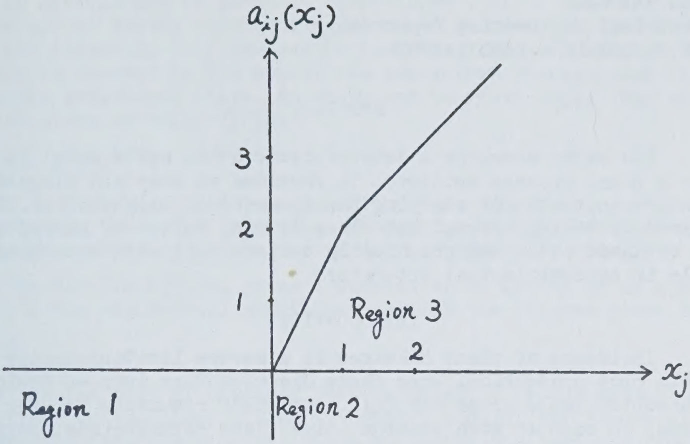

$a_{ij}(x_j)$
3
2
1
Region 3
1
2
x_j
Region 1
Region 2

Fig. 2

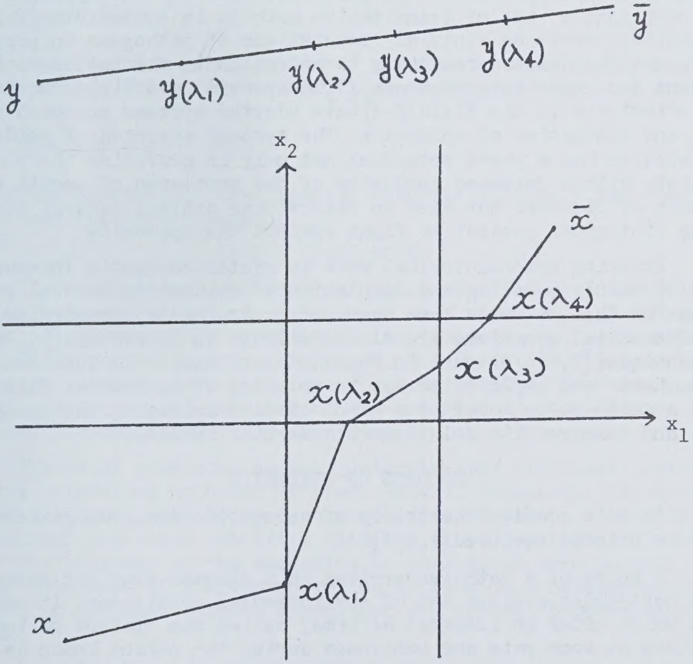

$\bar{y}$
$y(\lambda_4)$
$y(\lambda_3)$
$y(\lambda_2)$
$y(\lambda_1)$
y
x_2
$\bar{x}$
$x(\lambda_4)$
$x(\lambda_3)$
$x(\lambda_2)$
x_1
$x(\lambda_1)$
x

Fig. 3

STATE SPACE MODELING OF A PLANT DISEASE EPIDEMIC

R.K. VARSHNEY
Electrical Engineering Department
P.A.U. Ludhiana 141004 INDIA

ABSTRACT

The paper develops a deterministic state space model in discrete form for a plant disease epidemic. It provides an easy and elegant quantitative approach for studying the dynamics of epiphytotics. The model is analysed by taking assumed but close to real values of parameters. The results obtained and shown graphically compare well with experimental data available in epidemiological literature.

INTRODUCTION

Incidence of plant diseases is a severe limiting factor for profitable crop production. When these diseases turn into epidemics and become uncontrollable crops may fail completely resulting in near famine conditions. To counter such eventualities plant pathologists, particularly epidemiologists amongst them, make attempts to find efficient methods for forecasting and control of epidemics. These forecasts enable the farmers to take effective control measures at appropriate times for reducing the severity of the epidemocs. Pesticide industry can also benefit from such forecasts in planning and implementing their production schedule.

Efficient forecasting of epidemics requires the indentification of the causative parameters of the plant diseases and the dynamics of disease propagation. Use of quantitative methods in epidemiological investigations involving study of 'populations of pathogens in populations of hosts, and the disease resulting therefrom under the influence of the environment and human interferences'[1]has proved rewarding. Recently, several scientists in the field[2-5]have adopted systems approach involving modeling and simulation of epidemics. The systems approach in epidemiological studies has a great potential not only in providing the plant pathologists with a deepened knowledge of the processes of growth and propagation of diseases but also in determining optimal control strategies including biological control to fight against the epidemics.

Existing epidemiological work in system-theoretic framework consists of mainly modeling and simulation of epidemics. Several growth functions for the epidemics have been proposed. Models, consisting of first order differential equations are also available in literature[6]. But,state space techniques[7,8], elegant in form and explicative in function, do not appear to have been employed so far in modeling of epidemics. This paper presents a state space model of a plant disease epidemic, discusses its features and compares its solution with earlier results.

PROGRESS OF EPIDEMICS

In this section the theory of occupation and propagation of plant diseases is briefly reviewed[1,6,9].

A spore of a pathogen arrives in a disease-free but susceptible crop-the host-under a given set of environmental conditions. it establishes a lesion which after an interval of time, called the 'latent period',begins to sporulate at some rate and continues during the period known as 'infectious' producing more spores. Each effective spore thus produced establishes a lesion which after its own latent period produces new effective spores. This leads to multiplication of spores resulting in the growth of the disease in epidemic form. The process of pustule opening resulting in sporulation is known as 'apparition'. After the infectious period, the spores become ineffective and cease to multiply, but leave the host material diseased. This process is known as 'removal'. The severity of the epidemic is denoted by the sum of the cumulative removals and the disease in the infectious stage. An early and heuristic model for an epidemic was given by Müller[10] as

$$D = A(cP)^{t/T} \tag{1}$$

where D, A, C, P, T and t denote rate of increase of disease per generation, initial inoculum, number of infectious propagules disseminated per lesion, probability of infection, incubation period and time respectively.

Van der Plank[6] has given a succession of models of an epidemic. Starting with the exponential growth concept for the disease given by

$$\frac{dy}{dt} = ry \tag{2}$$

where y is the fraction of diseased host material, r is the infection rate and t the time, on the lines of Pearl Veruhlst model[11] of growth of an isolated species in its ecological niche suggested the logistic growth, including the boundedness of the susceptible host material.

$$\frac{dy}{dt} = ry(1 - y) \tag{3}$$

Van der Plank bases his infection rate on the fraction of the diseased material that has passed the latent period and puts

$$\frac{dy}{dt} = ry(1 - y) \tag{4}$$

where y is the fraction of the diseased material at time t that has passed the latent period. The model is modified by taking into account the removals

$$\frac{dy}{dt} = r(y - y)(1 - y) \tag{5}$$

where y is the fraction of the diseased material at time t that has passed the latent and infectious periods.

STATE SPACE MODEL

In formulating a mathematical model of a process, several simplifying but reasonable assumptions are made. In the present formulation, variations of environmental conditions like temperature, leaf wetness, humidity, light and wind gust are ignored. The duration of latent and infectious periods and infection rate are considered as constant parameters. Effects of parameter perturbations on the progress of the epidemic can be investigated through sensitivity analysis[12].

The model presented here is deterministic since most epiphytotic effects are determined uniquely by known causes. Stochastic features, if considered necessary, may be incorporated later. Further, since the epidemiological processes exhibit a distinct diurnal pattern, discrete model for the progress of the epidemics is considered more natural than its continuous counterpart. The incorporation of time lags between occupation and apparition and also between apparition and removal becomes simple and meaningful.

Let time be reckoned from the instant of occupation of the vacant site in the host material by an effective spore. If apparition takes place at the instant t_1 and removal at t_2 the latent and the infectious periods are the open time intervals $[0,t_1)$ and $[t_1,t_2)$ respectively. Denote the fraction of the host material occupied by the pathogens during the latent period of the disease at the instant k by $x_1(k)$. We will call this latent disease. Similarly, denote the fraction of the host material in the infectious stage at the instant k by $x_2(k)$. We will call this apparent disease. Further, denote the fraction of the host material that has passed the infectious stage at the instant k by $x_3(k)$. This may be called inactive disease. The state of the disease at the instant k may then be described by the vector

$$x(k) = \left[x_1(k) \quad x_2(k) \quad x_3(k)\right]' \tag{6}$$

where $0 \leq x_i(k) \leq 1, \quad i = 1,2,3; \quad 0 \leq x_1(k)+x_2(k)+x_3(k) \leq 1$

and $'$ denotes transpose.

In developing the model, it is assumed that after the establishment of initial inoculum, the latent disease is produced by apparent disease only. This assumption is similar to the concept of corrected infection rate given in[5] by Van der Plank. Also assume that the supply of host material to the pathogens for occupation is limited. Under these assumptions, it is proposed that if g(k) represents the fraction of the host material turned into latent disease because of sporulation during the interval of time from the instant k to k+1 and r is the basic infection rate per interval of time

$$g(k) = r\, x_2(k)\left(1 - x_1(k) - x_2(k) - x_3(k)\right) \tag{7}$$

for all k except when $x_1(k) + x_2(k) + x_3(k) > 1$

then for the lowest value of such k

$$g(k-1) = 1 - x_1(k-1) - x_2(k-1) - x_3(k-1) \tag{8}$$

Condition (8) imposed on g(k) is necessary in view of the fact that for no k, $x_1(k) + x_2(k) + x_3(k) \ngtr 1$

Now, if the integer p denotes the duration of latent period and x_o the initial value of the latent disease

$$x_1(k) = x_o \qquad 0 \leq k \leq p-1 \tag{9a}$$

$$x_1(p) = 0, \tag{9b}$$

since the latent disease will remain at its initial value during its latent period, will turn apparent at its end and vanish.

Later, at all the discrete instants of time the value of latent disease will be given by

$$x_1(k+1) = x_1(k) + g(k) - g(k-p), \qquad k \geqslant p \tag{9c}$$

the last term on the right hand side denotes the latent disease that has passed its period of latency and has become apparent.

As far as the apparent disease is concerned, during the first cycle of the latent period it does not exist. It appears immediately after it and retains its value during the infectious period denoted by i. Hence,

$$x_2(k) = 0 \qquad 0 \leq k \leq p-1 \tag{10a}$$

$$x_2(p+j) = x_o \qquad 0 \leq j \leq i-1 \tag{10b}$$

Further, after the first cycle of latent and infectious periods, it turns inactive and vanishes. This is expressed by

$$x_2(p+i) = 0 \tag{10c}$$

Now take up the case of apparent disease after the first cycle of latent and infectious periods. The apparent disease then is that latent disease which has passed its period of latency minus that part of the disease that has become inactive after passing through the infectious period. Using the principle of balance we have

$$x_2(k+1) = x_2(k) + g(k-p) - g(h-p-i), \qquad k \geqslant p+i \tag{10d}$$

As for the inactive disease, $x_3(k)$, which is known as cumulative removals in epidemiological literature, using similar arguments we have

$$x_3(k) = 0 \qquad 0 \leqslant k \leqslant p+i-1 \tag{11a}$$

$$x_3(p+i) = x_o \tag{11b}$$

during the first cycle of latent and infectious periods. Later,

$$x_3(k+1) = x_3(k) + g(k-p-i), \qquad k \geqslant p+i$$

If the severity of the disease at an instant k is denoted by $z(k)$, we have

$$z(k) = x_2(k) + x_3(k) \tag{12}$$

Equations (9)-(11) constitute the state model of an epiphytotic characterising the fractions of disease in its different stages - latent, apparent and inactive. The severity of the epidemic at the instant k is given by $z(k)$. The block diagram in fig.1 represents the model after the first wave of infection (for $k \geqslant p+i$). D_j and F in the diagram represent delay by j days and function generation activity.

EXAMPLE

Results are obtained for epidemics with parameters assumed by Zadoks[2] are taken for convenience of comparison. These parameters approximate the range prevalent for potato late blight(*Phytophthora infestans*) and wheat rusts (Puccinia graminis, P. recondita and P. Striiformis) The time interval between the instants k and k+1 has been chosen to be a day. Further, to provide an illustration the values of the parameters are taken as

Basic daily infection rate,	$r = 16$
latent period,	$p = 4$ days
infectious period,	$i = 1$ day
initial inoculum,	$x_o = 10^{-4}$

The results in respect of latent disease x_1, apparent disease x_2, inactive disease x_3 and the severity of disease y are shown in Figs. 2-5. The solution of the proposed state space model yields very meaningful results which compare very well with experimental data and are in agreement with theory[2].

The latent and the apparent diseases exist in pulses throughout the epidemic period but are able to produce a periodically rising staircase wave for the severity of disease. The time profile of the cumulative removals is also rising with time and lags the severity of disease curve by one day, the infectious period. The entire host material can be observed to be diseased on the 24th day from the establishment of the initial inoculum whereas the complete removal takes place on the 25th day.

More interesting results are obtained by varying the durations of latent and infectious periods as also the initial value of the inovulum.

CONCLUSION

The state space characterization of the plant disease epidemic allows the inclusion of the three distinct stages of disease viz. latent, apparent and inactive to be incorporated explicitly. The choice of discrete formulation makes the model simple, easy to tackle and meaningful. Delays between occupation, apparition and removal are included neatly. Parameter variations during the progress of the epidemic can also be handled without much difficulty.

The solution of the model for an assumed epidemic compares well with experimental data collected by epidemiologists and is in complete agreement with theory.

REFERENCES

1 Kranz, J., Epidemiology, Concepts and Scope in Current Trends in Plant Pathology, ed. S.P. Ray Chaudhuri & J.P. Verma, Lucknow, Deptt of Botany, Univ. of Lucknow, 1973.

2 Zadoks, J.C., System Analysis and the Dynamics of Epidemics, Phytophatology, Vol. 61, pp 600-610, 1971.

3 Caswell, H.C., Koenig H. E., Resh J. A., Ross, Q.E., An Introduction to Systems Science for Ecologists in System Analysis and Simulation in Ecology, ed. B.C. Patten, Vol II, pp 3-78, New York, Academic Press, 1972.

4 Kiyosawa, S., Mathematical Studies on the Curve of Disease Increase, Ann. Phytopath. Soc. Japan, Vol 38, pp 30-40, 1972.

5 Waggener, P.E., Horsfall, J. G., EPIDEM, A Simulator of Plant Disease written for a Computer, Conn. Agr. Exp. Stat. Bull. 698, 1969.

6 Van der Plank, J.E., Plant Diseases: Epidemics and Control, N.Y. Academic Press, 1963.

7 Timothy, L.K., Bona B.E. State Space Analysis: An Introduction, New York, McGraw Hill Book Co. 1968.

8 DeRusso, P.M., Roy, R.J., Close, C.M., State Variables for Engineers, New York, John Wiley, 1965.

9 Kranz, J. Epidemics of Plant Diseases Mathematical Analysis and Modelling, New York, Springer-Verlag, 1974.

10 Müller, K.O., Theoretische Betrachtungen zur Epidemiologic Avenochorischer Pflanzenkrankheiten, Z. Pflanzen Krankh, Vol 64, pp 402-407, 1957.

11 Rosen, R., Foundations of Mathematical Biology, Vol. III, New York, Academic Press, 1973.

12 Varshney, R. K., Reduced Order Sensitivity Models of Linear Systems with Applications, CSL Report No: R 545, Urbana, Coordinated Science Laboratory, Univ. of Illinois, 1972.

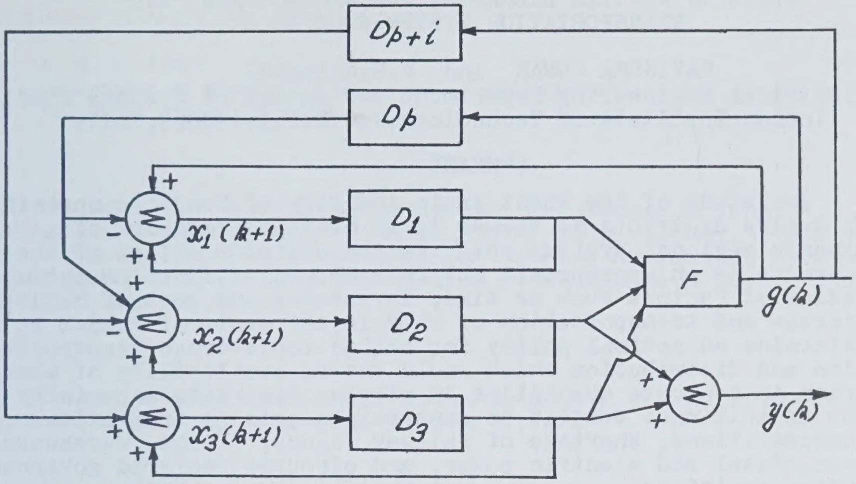

D_{p+i}
D_p
+
−
D_1
$x_1(k+1)$
+
+
F
$g(k)$
−
D_2
$x_2(k+1)$
+
+
+
$y(k)$
+
D_3
$x_3(k+1)$
+

Fig.1 Model of disease propagation after the first cycle of latent and infectious periods.

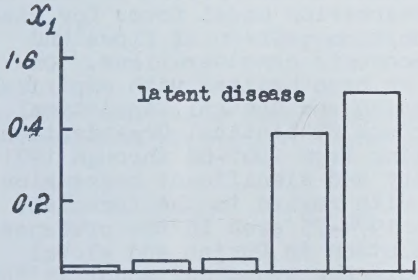


Fig. 2

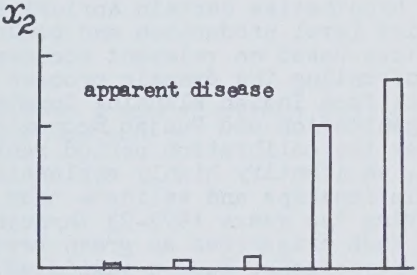


Fig. 3

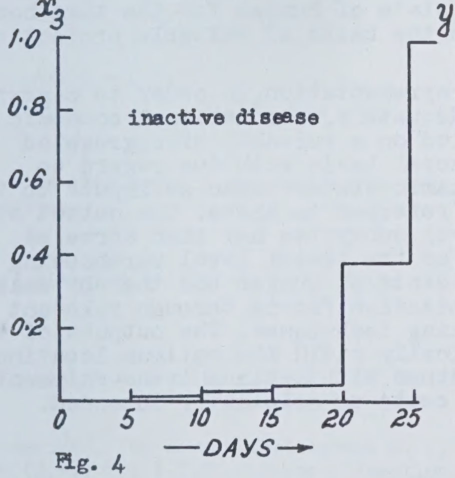


Fig. 4

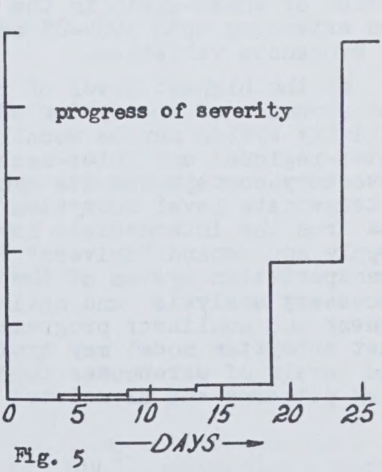


Fig. 5

PROCESS DYNAMICS OF WHEAT GRAIN SECTOR IN PUNJAB STATE IN A3-TIER HIERARCHY FOR WAREHOUSING AND TRANSPORTATION SYSTEM STUDIES

RAVINDRA KUMAR and P.S.SATSANGI
Electrical Engineering Department and School of Systems Studies
Indian Institute of Technology,New Delhi-110029,India

SUMMARY

The study of the wheat grain industry of Punjab comprising of twelve districts is viewed in an overall framework of a 3-tier dynamic regional systems analysis.The ultimate object of the exercise is an appropriate multiple objective function optimization of factors such as time, investment and cost of handling storage and transportation of wheat in the state of Punjab to determine an optimal policy for collection,storage,transportation and distribution which would ensure availability of wheat grain in adequate quantities in all the districts especially the deficit ones subject to constraints arising from price considerations, shortage of railway wagons, trucks, warehouses, coal,diesel and electric power, and ofcourse declared government policy decisions.

In the present paper we pick up at the intermediate level of the 3-tier heirarchy and model the process dynamics of the Punjab State with regard to its wheat industry.Towards this end, we hypothesize certain apriori regression model forms for district level production and consumption pattern of flows and prices based on relevant socio-economic considerations. On confronting the dynamic process so hypothesized with empirical data from Indian Planning Commission and Central Statistical Organisation and Punjab Economic and Statistical Organisation, over the calibration period ranging from 1961-62 through 1971-72, we identify highly explanatory and significant regression relationships and validate them with regard to the forecasts during the years 1972-73 through 1974-75 even in the presence of such exigencies as green revolution in Punjab and global energy crisis. Such a validated process dynamical model is then used for prediction of production and consumption flows and prices of wheat-grain in the state of Punjab for the time horizon extending upto 1984-85 on the basis of suitable projections of exogenous variables.

At the highest level of representation,in order to capture the dynamics of the system adequately, the national economic activity system may be modelled on a suitably disaggregated inter-regional and inter-sectoral basis with due regard to inventory concept and its dynamic states taken as inputs to the intermediate level subsystem referred to above. The output states from the intermediate level subsystem may then serve as supply and demand "drivers" for the lowest level warehousing and transportation system of the state of Punjab and thereby **enable necessary analysis** and optimization for it through relevant linear and nonlinear programming techniques. The outputs of this last subsystem model may typically yield the optimum locations and levels of warehouses together with optimum trans-shipment flow patterns and associated costs as originally intended.

ON INTEGRO-DIFFERENTIAL EQUATIONS ARISING IN POPULATION DYNAMICS AND CONTROL SYSTEMS

A. S. C. SINHA
Department of Electrical Engineering and Mathematics
IIT, Fort Wayne, IN 46803

ABSTRACT

The problem of stability properties for the solutions of Integro-differential systems often occurring in control theory, biology, economics, etc., is considered. Asymptotic stability criterion for functional differential equations is derived. The criterion is a condition on the roots of a certain "quasi-polynomial."

Several retarded differential systems in physics, control theory biology and economics can be given in the form [1]

$$\dot{x}(t) = Ax(t) + \sum_{j=1}^{m} \int_{-T}^{0} d\, n_j(t,\theta)\, x_j(t+\theta) \tag{1}$$

In previous works, systems of the form

$$\frac{dx(t)}{dt} = Ax(t) + \int_{-T}^{0} d\, n(\theta)\, x(t+\theta) \tag{2}$$

have been analyzed by reducing an equivalent system of ordinary differential equations with the operator on the right-hand side [2]

$$\frac{dx_t(\theta)}{dt} = Ax_t(\theta) \tag{3}$$

where $x_t(\theta) = x(t+\theta)$, $0 \geq \theta \geq -T$, and the operator A is defined as follows:

$$Ax(\theta) = \begin{cases} \dfrac{dx(\theta)}{d\theta}, & -T \leq \theta < 0 \\[2ex] \displaystyle\int_{-T}^{0} d\, n(\theta)\, x_j(t+\theta), & \theta = 0. \end{cases}$$

In this paper a new method for the generalized functional differential equations

$$\dot{x}(t) = A(t)\, x(t) + \sum_{j=1}^{m} \int_{0}^{t} K_j(t, s, x(s))\, ds \tag{4}$$

and

$$\dot{x}(t) = A(t)\, x(t) + \sum_{j=1}^{m} \int_{-h_j}^{0} K_j(t, \theta)\, x(t+\theta)\, d\theta \tag{5}$$

is derived. We assume that kernel K_j $(j = 1,\dots,m)$ is continuous for $t \in (T,\infty)$, $\theta \in [-T,0]$. Given a function $x_0 \in C[0,T]$, we say that a function x is a solution of (4) satisfying the initial condition

$$\forall\, t \in [0,T] : x(t) = x_0(t) \tag{6}$$

if (6) holds and (5) is true for $t \in [T,\infty)$.

Dunkel [3] studied equations of the form

$$\dot{N}(t) = \left[b + \int_{\tau}^{T} \psi(N(t - a))\, ds(a)\right] N(t) \tag{7}$$

which generalizes population models by Goel et al. [4], Sinha [5], Pielon [6], among many others.

The generalizations to the other types of population models given by Goel et al. and Pielon are obvious. Let N(t) and M(t) be the number of species of types A and B respectively at a time t. Additional terms P(•), Q(•) allow more interaction between species A and B. The equations are

$$\dot{N}_i(t) = N_i\left[A_i + \sum B_{ij} N_j + \int_0^r P(\theta)\, M_i(t - \theta)d\theta\right]$$

$$\dot{M}_i(t) = M_i\left[A_i' + \sum B_{ij}' M_j + \int_0^r Q(\theta)\, N_i(t - \theta)d\theta\right] \tag{8}$$

The additional interaction may, for example, be considered when species A has an unlimited food supply and species B relies upon A for his development. Some results of this model are given by May [7]. Since very complex equations describe the study of interacting models, where the present mathematical theory may be insufficient to deal with it, one is forced to linearize the model equation and study the model via simulation using experimental data.

Theorem 1: Suppose that the fundamental matrix ψ of $\dot{x}(t) = A(t)$ satisfies

(i) $|\psi(t)\, \psi^{-1}(\tau)| \leq \xi \exp[-\sigma(t - \tau)]$

for all $t, \tau \geq 0$, $\xi > 0$, $\sigma > 0$.

(ii) $\left|\int_0^t K_i(t, s, x(s))ds\right| \leq \mu_i \sup_{0 \leq s \leq t} |x(s)|$

for any function x(s) continuous on $0 \leq s \leq t$ such that $|x(s)| \leq \rho$ on this interval.

(iii) $\left|\sum_{i=1}^{m} \mu_i\, h_i(t)\, |Q(t)|\, e^{-\lambda h_i(t)}\right| < 1$

Then the solution of (4) approaches the limit

$$\lim_{t\to\infty} [x(t)\, e^{-\lambda t}] = \left\{1 + \sum_{i=1}^{m} \mu_i\, h_i(t)\, |Q(t)|\, e^{-\lambda h_i(t)}\right\}^{-1}$$

$$\cdot \left[\phi(0) + \sum_{i=1}^{m} e^{-\lambda h_i(t)} \int_{-h_i}^{0} e^{-\lambda s}\, \phi(s)\, ds\right]$$

exponentially under certain initial functions.

Lemma: [Hale, 74, 8] If there is a $\mu > 0$ such that $r(t) = \exp(-\mu t)$ for $t > 0$ and

$$\{\lambda \in C:\ \operatorname{Re}\lambda \geq 0,\ \det[\lambda I - L(e^{\lambda} \cdot I)] = 0\}$$

$$= \phi$$

then there are $\nu > 0$, $M > 0$ such that $|T(t)\phi| \leq Me^{-\nu t}\, |\phi|$.

Hale (1974) generalizes the results of Wright [9] and Hayes [10] who considered, respectively, the transcendental equation

$$\lambda e^{\lambda} + a = 0 \tag{9}$$

and

$$\lambda e + a\, e^{\lambda} + c = 0 \tag{10}$$

with the real coefficients a and c.

It should be pointed out that general results are known for locating zeros of functions of the form $P(\lambda, e^{\lambda})$, where $P(x, y)$ is a polynomial in x, y. For example, in a paper by Pontryagin [11], necessary and sufficient conditions are given for zeros of P to have real negative parts. As an example of this, see Bellman and Cooke [12] for a derivation of Hayes' results using the theorems of Pontryagin. We now consider the general case in which the coefficients and retardations in (4) and (5) are non-constant and result in the following theorem. The criterion is a condition on the roots of certain transcendental equations.

Theorem 2: If for each $t > t_0$

$$D(\lambda,t) \overset{\text{def}}{=} \det\left\{A_0^*(t,\lambda) - \sum_{j=1}^{m} \left[\left\|\int_{-T_j}^{0} e^{\lambda\theta}\, K_j(t,\theta)\, d\theta\right\|\right]\right.$$

$$\left. \times \left[\left\|\int_{-T_j}^{0} e^{\lambda\theta}\, K^T(t,\theta)\, d\theta\right\|\right]\right\} = 0$$

where

$$A^*(t,\lambda) = (2\lambda - m)\, I_n - (A(t) + A^T(t))$$

has real roots, let $\Lambda(t)$ denote the largest of these roots then if $\Lambda(t) \leq \delta < 0$, the trivial solution of (5) is asymptotically stable.

Remark: Under appropriate conditions, solutions of (5) result in studying the roots of transcendental equations where Newton's method of successive approximation can be applied to study the location of roots.

Proof of Theorem 1.

According to a previous result, if the solution of

$$\dot{x}(t) = A(t)\ x(t) \tag{11}$$

satisfies condition (i), then the following quadratic functional

$$V(t,\ \phi(\cdot)) = \langle Q(t)\ \phi(\cdot),\ \phi(\cdot)\rangle = \int_t^\infty |\psi(\tau)\ \psi^{-1}(t)\ \phi(\cdot)|^2\ d\tau \tag{12}$$

with some positive definite matrix Q(t) is a Liapunov function such that along the trajectories of (11)

$$\dot{V}(t,x(\cdot)) = \langle [A^T(t)\ Q(t) + Q(t)\ A(t) + \dot{Q}(t)]\ x(t),\ x(t)\rangle = -|x(t)|^2 \tag{13}$$

Computing $\dot{V}$ along the trajectories of (4) yields

$$\begin{aligned}\dot{V}(t,x(\cdot)) &= \langle [A^T(t)\ Q(t) + Q(t)\ A(t) + \dot{Q}(t)]\ x(t),\ x(t)\rangle \\ &\quad + 2\left(\sum_{j=1}^{m} \int_0^t K_j(t,\ s,\ x(s))ds,\ Q\,x(t)\right) \\ &= -|x(t)|^2 + 2\left(\sum_{j=1}^{m} \int_0^t K_j(t,\ s,\ x(s)),\ Q\ x(t)\right)\end{aligned} \tag{14}$$

we define $|x_t| = \mathrm{Sup}_{0<s<t}\ |x(s)|$.

If $Q = (q_{ij})$ and $|Q| = \left(\sum_{ij} q_{ij}^2\right)^{\frac{1}{2}}$, we obtain, using (ii),

$$\begin{aligned}&\left(\int_0^t K_j(t,\ s,\ x(s))ds,\ Q\ x(t)\right) \\ &\quad \le |Q| \left(\int_0^t K_j(t,\ s,\ x(s))ds\right) |x(t)| \\ &\quad \le |Q|\ |\mu_j|\ |x_t|\ |x| \\ &\quad \le \frac{|Q|\ |\mu_j|}{\Delta} \sqrt{V}(t,\ x(t)),\ \sqrt{V}(t,\ x_t(t))\end{aligned} \tag{15}$$

Since $\eta|x(t)|^2 \ge V(t,\ x(t)) \ge \Delta|x(t)|^2$. Thus, from (14), we have

$$\dot{v}(t) \equiv \dot{V}(t,\ x(\cdot)) \leq -|x(t)|^2 + \sum_{j=1}^{m} \frac{|Q|\ |\mu_j|}{\Delta} \sqrt{V}(t,\ x(t))\ \sqrt{V}(t,\ x_t) \tag{16}$$

$$\leq -2\delta\ v(t) + 2 \sum_{j=1}^{m} \frac{|Q|\ |\mu_j|}{\Delta} \sqrt{v}(t)\ \sqrt{v}_t \tag{17}$$

where $\delta = \frac{1}{2\eta}$, $v(t) \equiv V(t,\ x(\cdot))$.

We shall now compare (17) with

$$\dot{u}(t) = -\delta\ u(t) + 2 \sum_{j=1}^{m} \frac{|Q|\ |\mu_j|}{\Delta} \sqrt{u}(t),\ \sqrt{u}_t \tag{18}$$

to obtain asymptotic correspondence between the two systems. We transform equation (18) by substituting $\omega(t) = \sqrt{u}(t)$ to obtain

$$\dot{\omega}(t) = -\delta\ \omega(t) + \sum_{j=1}^{m} \frac{|Q|\ |\omega_j|}{\Delta} \omega(t - h_i),\quad h_i > 0. \tag{19}$$

Again substituting $z(t) = e^{-\lambda t}\ \omega(t)$ in (19), we have

$$\dot{z}(t) = -\sum_{j=1}^{m} \frac{|Q|\ |\mu_j|}{\Delta} e^{-\lambda h_j} [z(t) - z(t - h_j)] \quad \text{for} \quad t \geq 0$$

$$z(t) = \phi(t)\ e^{-\lambda t} \quad \text{for} \quad -h_j \leq t \leq 0 \quad \text{for each } j. \tag{20}$$

These equations are then equivalent to

$$\dot{z}(t) = -\sum_{j=1}^{m} \frac{|Q|\ |\mu_j|}{\Delta} e^{-\lambda h_j} \int_{t-h_j}^{t} z(s)ds + c;\quad t \geq 0$$

$$c = \phi(0) + \sum_{j=1}^{m} \frac{|Q|\ |\mu_j|}{\Delta} e^{-\lambda h_j} \int_{-h_j}^{0} \phi(s)ds \tag{21}$$

Equation (21) may be translated to make it homogeneous by a substitution

$$\left(1 + \sum_{j=1}^{m} \frac{|Q|\ |\mu_j|}{\Delta} h_j\ e^{-\lambda h_j}\right) y(t) = \left(1 + \sum_{j=1}^{m} \frac{|Q|\ |\mu_j|}{\Delta} h_j\ e^{-\lambda h_j}\right) z(t) - c$$

where c is defined earlier. The equation (21) then yields

$$y(t) = -\sum_{j=1}^{m} \frac{|Q|\ |\mu_j|}{\Delta} e^{-\lambda h_j} \int_{t-h_j}^{t} y(s)ds;\quad t \geq 0$$

with initial function

$$y(t) = \phi(t)\ e^{-\lambda t} - \frac{c}{\left(1 + \sum_{j=1}^{m} \frac{|Q|\ |\mu_j|}{\Delta} h_j\ e^{-\lambda t}\right)} \quad \text{for } -h_j \leq t \leq 0 \tag{22}$$

Estimating with respect to the norm yields

$$\|y(t)\| = \sum_{j=1}^{m} \frac{|Q|\ |\mu_j|}{\Delta}\ e^{-\lambda h_j} \int_{t-h_j}^{t} \|y(s)\|\, ds$$

Let $\|y(t)\| \leq M$ on $-h_j \leq t \leq 0$, then it follows that $\|y(t)\| \leq M$ for all $t \geq 0$. Assume $\|y(t)\| \nleq M$ for all $t \geq 0$, then given $\varepsilon > 0$, there is a $t^* > 0$ such that $\|y(t)\| < M + \varepsilon$ for $-h < t < t^*$. Let t_1^* be the first point at which $\|y(t_1^*)\| = M + \varepsilon$. Therefore,

$$\|y(t_1^*)\| \leq \sum_{j=1}^{m} \frac{|Q|\ |\mu_j|}{\Delta}\ e^{-\lambda h_j} (M + \varepsilon)\ h_j$$

From assumption (iii), it follows that at $t = t_1^*$,

$$\|y(t_1^*)\| = (M + \varepsilon) \leq \left\| \sum_{i=1}^{m} |Q|\ |\mu_i|\ h_i\ e^{-\lambda h_i} \right\| < M + \varepsilon$$

which is in contradiction to our hypothesis, implying that $\|y(t)\| \leq M$ for all $t \geq 0$. Next, to show that $\|y(t)\| \to 0$ exponentially, we use induction arguments. Since $\|y(t)\| \leq M$ for $t \geq 0$, we show that such a solution satisfies an estimate

$$\|y(t)\| \leq \left\| \sum_{i=1}^{m} |Q|\ |\mu_i|\ h_i\ e^{-\lambda h_i} \right\|^{n} M$$

for all $t \geq t_0 + nt_0$, $n = 0, 1, \ldots, t_0 \geq 0$. By standard induction argument, the assertion follows. Since

$$\left\| \sum_{i=1}^{m} |Q|\ |\mu_i|\ h_i\ e^{-\lambda h_i} \right\| < 1$$

implies that $\|y(t)\| \to 0$ as $\eta \to \infty$. The asymptotic stability of $x(t)$ is then claimed by standard comparison theorems.

Proof of Theorem 2. Let $y_i(t) = e^{-\lambda t} x_i(t)$, $i = 1, \ldots, n$, where λ is a fixed real number. Substituting into (5), we have

$$\dot{y}(t) = (A(t) - \lambda I)\ y(t) + \sum_{\ell=1}^{m} \int_{-h_\ell}^{0} e^{\lambda\theta},\ K_\ell(t,\theta)\ y(t+\theta)\, d\theta \tag{23}$$

We will now investigate the stability of the trivial solution. To this end, we define a Liapunov functional

$$v(t) = \sum_{i=1}^{n} y_i^2(t) + \sum_{i=1}^{n} \sum_{\ell=1}^{m} \int_{t-h_\ell}^{t} y_i^2(s)\,ds \tag{24}$$

It is easily seen that this functional satisfies the conditions of Hale's stability theorem [8]. Along the trajectories of (5), we have

$$\begin{aligned} -\dot{v}(t) \leq & \sum_{i=1}^{n} \sum_{j=1}^{m} \Big\{(2\lambda - m)\,\delta_{ij} - 2A_{ij}(t)\Big\}\, y_i(t)\, y_j(t) \\ & + \sum_{i=1}^{n} \sum_{\ell=1}^{m} y_i^2(t - h_\ell) \\ & - 2\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{\ell=1}^{m} \int_{t-h_\ell}^{t} e^{\lambda\theta}\, K_{ij}(t,\theta)\,d\theta\; y_i(t)\; y_j(\cdot) \end{aligned} \tag{25}$$

where define $|y_t| = \mathrm{Sup}_{0<s<t}\, |y(s)|$

The matrix associated with the quadratic form (25) can be written in the block matrices as

$$-D(\lambda,t) = \begin{bmatrix} A^*(t,\lambda),\; -\Big|\Big|\int_{t-h_1}^{t} e^{\lambda\theta}\, K_1(t,\theta)\,d\theta\Big|,\; \dots,\; -\Big|\Big|\int_{t-h_m}^{t} e^{\lambda\theta}\, K_m(t,\theta)\,d\theta\Big| \\ -\Big|\Big|\int_{t-h_1}^{t} e^{\lambda\theta}\, K_1^T(t,\theta)\,d\theta\Big|,\; 1,\; 0,\; \dots,\; \dots,\; 0 \\ \cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot \\ -\Big|\Big|\int_{t-h_m}^{t} e^{\lambda\theta}\, K_m(t,\theta)\,d\theta\Big|,\; 0,\; \dots,\; 0,\; 1 \end{bmatrix}$$

where

$$A^*(t,\lambda) = (2\lambda - m)I_n - (A_{ij}(t) + A_{ji}(t))$$

For sufficiently large λ, the roots of $|D(\lambda,t) - \mu I| = 0$ shall be real and positive. For sufficiently small λ, the quadratic form shall be non-positive and since the roots of $|D(\lambda,t) - \mu I| = 0$ are continuous functions of λ, at least one root shall be zero for some value of λ. These values of λ are the values of λ satisfying $|D(\lambda,t)| = 0$. It is easily seen that

$$\begin{aligned} |D(\lambda,t)| = \Bigg| A_0^*(t,\lambda) - \sum_{j=1}^{m} \Bigg|\Bigg|\left[\int_{-T_j}^{0} e^{\lambda\theta}\, K_j(t,\theta)\,d\theta\right]\Bigg|\Bigg| \\ \times \Bigg|\Bigg|\left[\int_{-T_j}^{0} e^{\lambda\theta}\, K_j^T(t,\theta)\,d\theta\right]\Bigg|\Bigg|\Bigg| = 0 \end{aligned}$$

Let $\Lambda(t)$ denote the largest invariant roots of $D(\lambda,t)$. Suppose $\Lambda(t) \leq \mu < 0$ and take $\mu < \lambda_0 < 0$, then the quadratic form shall be positive definite for all $t < t_0$ if $\lambda = \lambda_0$. This implies that $y_i(t) = \exp(|\lambda_0|\ t)\ x_i(t) \to 0$ as $t \to \infty$ which proves the asymptotic stability.

REFERENCES

1. Hale, J.; Functional Differential Equations, Springer-Verlag, New York, 1971.

2. Markushin, E. M. and Shimanov, S. N.; Approximate Analytic Design of Controls for Systems with Time Lag, Avtomatika 1 Telemekhanika, No. 3, pp 13-20, 1968.

3. Dunkel, G. M.; Some Mathematical Models for Population Growth with Lags, Tech. Note BN-548. The Institute for Fluid Dynamics and Applied Mathematics, University of Maryland, 1968.

4. Goel, N. S., Maitra, S. C. and Montroll, E. W.; On the Volterra and Other Nonlinear Models of Interacting Populations, Rev. Mod. Physics, 43, 231-276, 1971.

5. Sinha, A. S. C.; Some Properties of Nonlinear Models of Volterra-Lotka Systems, 1974 IEEE Conference on Decision and Control including 13th Symposium on Adaptive Process, November 20-22, 1974, Phoenix, Arizona.

6. Pielow, E. C.; An Introduction to Mathematical Ecology, Wiley-Interscience, New York, 1969.

7. May, R. M.; Time-Delay Versus Stability in Population Models with Two and Three Trophic Levels, Ecology, Vol. 54, No. 2, 1973, pp 315-325.

8. Hale, J. K.; Functional Differential Equations with Infinite Delays, J. of Math. Anal. and Applications, 48, 276-283, 1974.

9. Wright, E. M.; Stability Criteria and the Real Roots of a Transcendental Equation, SIAM J. Appl. Math. 9, 1961, pp 136-148.

10. Hayes, N. D.; Roots of Transcendental Equations Associated with a Certain Difference-Differential Equation, J. London Math. Soc., 25, 1950, 226-232.

11. Pontryagin, L. S.; On the Zeros of Some Elementary Transcendental Functions, Amer. Math. Soc. Trans. (2), 1955, pp 95-110.

12. Cooke, K. L.; Linear Functional Differential Equations of Asymptotically Autonomous Type, J. Diff. Equations, 1970, 7, pp 154-174.

SOCIO-ECONOMIC SYSTEMS MODELLING AND ANALYSIS OF A REAL-LIFE COMMUNITY HEALTH CARE SYSTEM

L.C.AGARWAL and P.S.SATSANGI
Department of Electrical Engineering &
School of Systems & Management Studies
Indian Institute of Technology, Delhi
New Delhi-110029, India

J.B.ELLIS
Faculty of Environmental Studies,
York University,
Downsview, Ontario
(Canada)

ABSTRACT

The paper applies the component-to-system modelling construct of physical system theory to formulate "relatively non-interacting" through and across system state-models for a community health care system. These models are empirically calibrated and validated for predictive use in the context of a real-life health care system represented by the I.I.T.Delhi Sick Bay.

INTRODUCTION

The theoretical structure of a community health care delivery system can be made identical to that of a physical system and rendered amenable to the physical system theoretic dynamic patient state model formulation and analysis which effectively add "new" dimensions of understanding and capability to the health care delivery problem. Community health is a dynamic concept which changes with time. The operation of a community health care system can be considered as a three staged process [1,2], the basic health needs perceived by the community, a few of which get converted into demands, which in turn are satisfied by resources or system facilities. The real life health service system represented by the Sick Bay at the I.I.T. Delhi, has served as the backdrop for the evolution of the system state models.

The health service system has been identified here as an aggregate of inter-connected components with each component modelled in terms of two complementary variables - a through variable Y, representing the flow of "patients" and an across variable X, representing their per unit values. The various health services aspects identified here as sub-systems or components (shown schematically in Fig.1) are: (1) health demand generation component G, (2) week day adjustment component W, (3) divertor component A, (4) diagnosis component H, (5) treatment and rehabilitation component T, (6) feedback component F, (7) outside treatment component E, (8) feedback component Fe, (9) distributor component B, and (10) control component C.

Individual mathematical component models are formulated with reference to component terminal graphs whose union gives rise to the system graph (shown schematically in Fig.2). Interconnection constraint equations in terms of fundamental cutset and circuit equations are written on the basis of this system graph. These equations are appropriately combined with the component terminal equations to develop "relatively non-interacting" through-variable and across variable system state-models for the health-care delivery system, such that they admit of a sequential solution in which we solve for the through variables first and the across variables next.

Before considering individual components in mathematical model terms we address ourselves to a discussion of certain related concepts, demographic projection and patient state concept, which have particular significance in our systems study.

2. RELATED CONCEPTS

2.1 Demographic Group Classification of the I.I.T.Population:

A lumped population model describing the time path of I.I.T. population does not contain adequate information to predict the health needs and demands accurately. Our demographic model, has been developed from the health systems point of view. Accordingly the I.I.T. population has been so classified as to have "homogeneous" groups of people who in a broad sense, have almost similar health need generation and need-to-demand conversion attitude towards health resources. While forming these groups, other factors such as age, sex, per capita income, educational level, way of living, diet and health lookout, residential and other facilities have also been considered. The I.I.T. Delhi population has been conceptually classified into four major groups as follows: (1) Male and female unmarried undergraduate, postgraduate and research students; (2) Faculty members, administrative officers and married M.Tech and research students, and their families; (3) supporting ministerial, technical and other class III employees and their families; (4) Office bearers, attendants and other class IV employees and their families.

All these groups have been further subdivided according to age and sex. For example three age groups have been considered in this study, namely 0-15 years, 16 to 50 years and above 50 years. Further sex-wise subdivision only in respect of middle age group (16 to 50 years) is considered relevant. Thus in all 13 sub-categories have been considered recognising that the entire male and female student population falls in one single sub-category.

2.2 Patient-State Concept:

Patient states are conceptualized such that each individual of the community under consideration may be identified with one of the four health states as given below. These health states are obviously homogeneous with respect to such features as health resource inputs and similar recovery periods among other things.

State 0: It consists of well individuals including all unreported cases*, whose number is taken to be Y_o.

State 1: It consists of individuals affected by minor injury or disease who report to the health centre and whose number is taken to be Y_1.

State 2: It consists of individuals affected by medium injury or disease who report to the centre and whose number is taken to be Y_2.

State 3: It consists of individuals affected by major injury or disease (serious cases) who report to the health centre and whose number is taken as Y_3.

---

*All individuals perceiving health needs,but not reporting at any of the health service centres are deemed as well individuals in this study.

In general, the patient state transition from time period t to t+1 is given by

$$Y(t+1)=T\,Y(t) \qquad (1)$$

where $Y = \left[Y_o, Y_1, Y_2, Y_3\right]^T$ and $T=\left[t_{ij}\right]$ such that in general $\sum_{i=0}^{3} t_{ij}=1$ for all j=0,1,2,3 and t= the sampling period (e.g, one day).

The transition matrix T, representing transition among patient health states from one time period to the next, is considered in the context of each of the three dynamic components of the system,viz, G,T and E.

3. COMPONENT MODELS

3.1 Health demand generation component G:

This is a dynamic component which describes the new health demand generation. It is hypothesized to be characterized by the schematic block diagram as in sub-fig.1.1, the terminal graph as in sub-fig.2.1, the discrete-time state equations as in equations (2a) and (2b), and the output equations as in equations (3a) and (3b).

$$S_g(t+1) = B_g\, Y'_g(t)+D_g\, R'_g(t) \qquad (2a)$$

$$\bar{S}_g(t+1) = B^*_g\, X'_g(t)+ D^*_g\, \bar{R}'_g(t) \qquad (2b)$$

$$Y''_g(t) = S_g(t)\ , \quad (3a)\ , \qquad X''_g(t) = -\,\bar{S}_g(t) \qquad (3b)$$

where, $S_g = \left[S_{og}, S_{1g}, S_{2g}, S_{3g}\right]^T$, a flow state-vector of through-variables.

$\bar{S}_g = \left[\bar{S}_{og}, \bar{S}_{1g}, \bar{S}_{2g}, \bar{S}_{3g}\right]^T$, a per-unit-value state-vector of across variables:

$Y'_g = \left[Y'_{og1}, Y'_{og2}, \ldots, Y'_{ogn}\right]^T$, an input flow vector.

$X'_g = \left[X'_{og1}, X'_{og2}, \ldots, X'_{ogn}\right]^T$, an input p.u. value-vector.

$Y''_g = \left[Y''_{og}, Y''_{1g}, Y''_{2g}, Y''_{3g}\right]^T$, an output flow vector.

$X''_g = \left[X''_{og}, X''_{1g}, X''_{2g}, X''_{3g}\right]^T$, an output p.u.value-vector.

$B_g = \left[b_{ij}\right]_g = \left[g_{ij}+f_{ij}(t)\right]_g$ and $B^*_g = \left[b_{ij}.Y'_{ogj}(t)/S_{ig}(t+1)\right]_g$ where

i=0,1,2,3 (patient health states), j=1,2,...,n(population groups)

$D_g = \left[d_{ij}\right]$ and $D^*_g = \left[d_{ij}.R'_{gj}(t)/S_{ig}(t+1)\right]$,

i = 0,1,2,3 and j=1,2,3,4 (resource inputs)

such that $\sum_{i=o}^{3} g_{ij} = 1$, $f_{oj}(t) = \sum_{i=1}^{3} f_{ij}(t)$ for all j=1,2,..n;

and $d_{oj} = -\sum_{i=1}^{3} d_{ij}$ for j=1,2,3,4

g_{ij}'s are "normal" health demand generation parameters, f_{ij}'s are epidemic dependent health demand generation parameters,and d_{ij}'s are resource-dependent health demand generation parameters.

At any time t, the new health demand generation in any patient state (effectuated through parameters b_{ij} and d_{ij}) depends upon such factors as population group characteristics, preventive measures, social characteristics of the campus environment, available health service resources, financial assistance policy, epidemic spread function and seasonal variations in population.

3.2 Week day adjustment component W:

This algebraic component is assumed to be described by the block diagram in sub-fig.1.2, terminal graph in sub-fig.2.2 and the terminal equations as under,

$$Y''_w(t) = W_w Y'_w(t) \quad (4a) \text{ and } X''_w(t) = -W_w^{*T} X'_w(t) \quad (4b)$$

where,

$W_w = [w_{ij}]_w$ and $w_w^{*T} = [w_{ij}^{*T}]_w$, i=0,1,2,3 and j=1,2,3,4.

Due to different patterns of working hours of the I.I.T.Sick Bay and outside treatment agency on weekdays, week-ends and holidays, this component provides necessary adjustment in patient turn-outs on weekdays (working Mondays through Fridays), week-ends (working Saturdays) and holidays(including Sundays).

3.3 Divertor component A:

Depending on individual/community knowledge and perception of relative attractiveness of sickbay and accessibility of other health facilities, the patients may either report at Sick Bay or go for outside treatment. Such a diversion of patients is conceptualized to be effectuated by this component, which is described by the block diagram in sub-fig.1.3, terminal graph in sub-fig.2.3 and the component terminal equations as follows;

$$Y''_a(t) = \alpha Y'_a(t) \text{ ; and} \quad (5a)$$

$$X''_a(t) = -X'_a(t) = X'''_a(t) \quad (5b)$$

where,

α is a diagonal matrix.

3.4 Diagonosis Component D:

This algebraic component is assumed to be described by the block diagram (sub-fig.1.4), terminal graph (sub-fig.2.4) and the terminal equations as under:

$$Y''_d(t) = H_d Y'_d(t) \quad (6a), \quad X''_d(t) = -H_d^{*T} X'_d(t) \quad (6b)$$

where, $H_d = [h_{ij}]_d$ and $H_d^{*T} = [h_{ij}^{*T}]$, $\begin{cases} i = 1dh, 2dh, 2dH, 3dH, 2d, 3d. \\ j = 1, 2, 3. \end{cases}$

The diagnosis component performs the function of medically advising the patients on the basis of examination either to proceed for treatment and rehabilitation within the Institute in health states 1,2 and 3 or to seek outside specialist treatment and rehabilitation in health states 2 and 3. In the former case, there are four streams, two of which correspond to treatment and rehabilitation at home in states 1 and 2 represented by 1dh and 2dh, while the other two correspond to treatment and rehabilitation in Institute Sick Bay in states 2 and 3 represented by 2dH and 3dH. In the latter case, there are two streams corresponding to outside specialist treatment and hospital rehabilitation in states 2 and 3 which are represented by 2d and 3d.

3.5 Treatment and rehabilitation component T:

This is hypothesized to be characterized by the schematic block diagram as in sub-fig.1.5, the terminal graph as in sub-fig.2.5, and the discrete-time state model as in equations (7) and (8):

$$S_t(t+1) = A_t\, S_t(t) + B_t\, Y'_t(t) + D_t\, R'_t(t) \qquad (7a)$$

$$\bar{S}_t(t+1) = A^*_t\, \bar{S}_t(t) + B^*_t\, x'_t(t) + D^*_t\, \bar{R}'_t(t) \qquad (7b)$$

$$Y''_t(t) = S_t(t) \quad \ldots.(8a), \qquad X''_t(t) = -\bar{S}_t(t) \qquad (8b)$$

where, $A_t = A_T(I-F_f)K_2$.

where, $S_t, \bar{S}_t$ have interpretations analogous to their counterparts for component G in section 3.1 and R'_t and $\bar{R}'_t$ are respectively input-resource flow-vector and input resource p.u. value-vector.

$$\left.\begin{array}{l} A_t = [a_{ij}]_t \quad \text{and} \\ A^*_t = [a_{ij} . S_{jg}(t)/S_{ig}(t+1)]_t \end{array}\right\} \quad i,j = 0,1,2,3$$

$A_T = [A_{ij}]_T$, $i=0,1,2,3$ & $j = 1,2,3$

I = unit matrix, F_f is defined in section 3.6 and K_2 in section 4.

$B_t = [b_{ij}]_t$ and $B^*_t = [b_{ij} .\, Y'_j(t)/S_{it}(t+1)]_t$

$i = 0,1,2,3$ and j = 1th, 2th, 2tH, 3tH

$D_t = [d_i]_t$ and $D^*_t = [d_i . R'_t(t)/S_{it}(t+1)]_t$

for all $i=0,1,2,3$ such that $\sum_i b_{ij}=1$ and $d_0 = -(d_1+d_2+d_3)$

b_{ij}'s are "normal" treatment and rehabilitation parameters, d_i's are resource dependent treatment and rehabilitation parameters.

3.6 Feedback component F:

This algebraic component is assumed to be described by the block diagram in sub-fig.1.6, terminal graph in sub-fig. 2.6 and the terminal equations as under:

$$Y''_f(t) = F_f\, Y'_f(t) \quad \ldots (9a), \qquad X''_f(t) = -F^{*T}_f\, X'_f(t) \qquad (9b)$$

where, $F_f = [f_{ij}]_f$ and $F^{*T}_f = [f^{*T}_{ij}]_f$, $i=0,1,2,$ & $j = 1,2,3$.

In order to capture the influence of different working hours of I.I.T.Sick bay on working Monday through Friday, working Saturday and Sunday and holidays of the week, this component regulates the output feedback appropriately.

3.7 Outside treatment component E:

This is a dynamic component which represents the overall outside diagnosis, treatment and rehabilitation facility, and may be characterized by the state model as in equation (10) & (11) with reference to the schematic block diagram (sub-fig. 1.7) and the terminal graph (sub-fig.2.7):

$$S_e(t+1) = A_e S_e(t) + B_e Y'_e(t) + D_e R'_e(t) \qquad (10a)$$

$$\bar{S}_e(t+1) = A^*_e \bar{S}_e(t) + B^*_e X'_e(t) + D^*_e \bar{R}'_e(t) \qquad (10b)$$

$$Y''_e(t) = S_e(t) \quad \ldots(11a), \qquad X''_e(t) = -\bar{S}_e(t) \qquad (11b)$$

where, $A_e = A_E(I-F_{fe})K_2$
where, $S_e, \bar{S}_e, A_e, A^*_e, D_e, D^*_e$ have interpretations analogous to their counterparts for components G and T in sections 3.1 &3.5,

$A_E = [A_{eij}]_E$, $i=0,1,2,3$ & $j = 1,2,3$.

$I =$ unit matrix, F_{fe} is defined in section 3.8 and K_2 in Section 4.

$B_e = [b_{ij}]_e$ and $B^*_e = [b_{ij} . Y'_{je}(t)/S_{ie}(t+1)]_e$

$i=0,1,2,3$ and $j=1,2,3$ such that $\sum_i b_{ij}=1$ and $d_o=-(d_1+d_2+d_3)$.

3.8 Feedback Component F_e:

This algebraic component is assumed to be described by the block diagram in sub-fig.1.8, terminal graph in sub-fig.2.8 and the terminal equation as under:

$Y''_{fe}(t) = F_{fe}\, Y'_f(t)$ ... (12a), $X''_{fe}(t) = -F^{*T}_{fe}\, X'_{fe}$ (12b)

where, $F_{fe} = [f_{ij}]_{fe}$ & $F^{*T}_{fe} = [f^{*T}_{ij}]_{fe}$, $i=0,1,2,$& $j=1,2,3$.

In order to reflect adequately the effects of working hours of outside treatment agency on first five days, sixth and seventh day of the week, this component moderates the output feedback appropriately.

3.9 Distributor component B:

The outgoing well individuals from each of the three dynamic components namely G, T and E, join the respective well population groups after distribution in this algebraic component which is described by the schematic block diagram as in sub-fig.1.9, the terminal graph as in sub-fig.2.9. and the terminal relations as in equations (13):

$Y''_b(t) = B_b\, Y'_b(t)$ ... (13a) and $X''_b(t) = -B^{*T}_b\, X'_b(t)$ (13b)

where, $B_b = [b_{ij}]_b$ and $B^{*T}_b = [b^{*T}_{ij}]_b$, $i=1,2,\ldots n$, & $j=1,2,3$.

3.10 Control Component C:

This special component consisting of all policy or control through and across variables in the system is hypothesized to be characterized by specified (through or across) drivers as follows:

$R = [R_{g1}, R_{g2}, R_{g3}, R_{g4}, R_t, R_e]^T$ is a specified through driver vector implied by light edges oriented away from G; T&E to 1,2,3,4; 5 & 6 respectively (Fig.2)

$\bar{R}' = [\bar{R}'_{g1}, \bar{R}'_{g2}, \bar{R}'_{g3}, \bar{R}'_{g4}, \bar{R}'_t, \bar{R}'_e]^T$ is a specified across deriver vector represented by heavy edges oriented towards G, T&E from 1,2,3,4; 5 & 6 respectively.

4. INTERCONNECTION CONSTRAINT EQUATIONS

The interconnection constraint equations in the form of the following fundamental cutset and fundamental circuit equations are written from the system graph in Fig.2.

F-Cutset equations: (Corresponding to formulation tree in heavy lines), $R'=R$, $Y'_w=Y''_g$, $Y'_a=K_2Y''_w$, $Y'_d=Y''_a+Y''_f$, $Y'_f=K_2Y''_t$; $Y'_t=K_7Y''_d$, $Y'_e=Y''_a+K_8Y''_d+Y''_{fe}$, $Y'_{fe}=K_2Y''_e$, $Y'_{obw}=Y''_{ow}$, $Y'_{obt}=Y''_{ot}$, $Y'_{obe}=Y''_{oe}$, $Y'_g=Y''_b+\Delta p$. (14a)

F-Circuit equations: (Corresponding to formulation tree in light lines), $R' = -R$, $X'_w = -X''_g$, $X'_a = -K_2 X''_w$, $X'_t = -K_7 X''_d$; $X'_d = X''_f - X''_a$, $X'_f = -K_2 X''_t$, $X'_e = -X''_a = -K_8 X''_d = -X''_{fe}$, $X'_{fe} = -K_1 X''_e$, $X'_{obt} = -K_3 X''_t$ or $-X''_{ot}$, $X'_{obe} = -K_3 X''_e$ or $-X''_{oe}$, $X'_{obw} = -K_3 X''_w$ or $-X''_{ow}$, $X'_g = -\Delta\bar{p} = -X''_b$. (14b)

where,

$$K_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad K_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}, \quad K_4 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T, \quad K_7 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$K_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad K_5 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad K_6 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad K_8 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

5. SYSTEM STATE MODEL

The foregoing constraint equations may be appropriately combined with the component terminal equations to develop the desired state space characterization representing the behaviour of the system as an aggregate of its components and interconnections[3].

As a simplifying feature notice that it is possible to formulate the system model in terms of separate through-variable and across-variable state-models, in as much as each of the three dynamic components G,T, and E are characterized in terms of relatively decoupled through-variable and across-variable state models.

Considering the direct sum of the through-variable and across-variable state equations for the three dynamic components G,T and E and expressing secondary variables[3]in terms of primary state and control variables through the help of appropriate f-cutset and f-circuit equations and algebraic through-variable and across variable component terminal equations, we obtain the system state models as follows:

Through Variable System Model:

$$\begin{bmatrix} S_g(t+1) \\ S_t(t+1) \\ S_e(t+1) \end{bmatrix} = \begin{bmatrix} (B_g B_b K_4 K_3 W_w), & (B_g B_b K_6 K_3), & (B_g B_b K_5 K_3) \\ (B_t K_7 H_d \alpha K_2 W_w), & \{(B_t K_7 H_d F_f K_2 + A_t)\}, & (0) \\ B_e\{(I - \alpha + K_8 H_d \alpha) K_2 W_w\}, & (B_e K_8 H_d F_f K_2), & \{B_e F_{fe} K_2 + A_e\} \end{bmatrix} \times \begin{bmatrix} S_g(t) \\ S_t(t) \\ S_e(t) \end{bmatrix} + \begin{bmatrix} D_g & 0 & 0 \\ 0 & D_t & 0 \\ 0 & 0 & D_e \end{bmatrix} \begin{bmatrix} R_g(t) \\ R_t(t) \\ R_e(t) \end{bmatrix} + \begin{bmatrix} B_g \Delta p(t) \\ 0 \\ 0 \end{bmatrix} \qquad (15a)$$

Across Variable System State Model:

$$\begin{bmatrix} \bar{S}_g(t+1) \\ \bar{S}(t+1) \\ \bar{S}_e(t+1) \end{bmatrix} = \begin{bmatrix} (B_g^* B_b^{*T} K_4 K_3 W_w^{*T}), & (B_g^* B_b^{*T} K_6 K_3), & (B_g^* B_b^{*T} K_5 K_3) \\ (0), & (A_t^* + B_t^* K_7 H_d^{*T} F_f^{*T} K_2), & (0) \\ (0), & (0), & (A_e^* + B_e^* F_{fe}^{*T} K_2) \end{bmatrix} \begin{bmatrix} \bar{S}_g(t) \\ \bar{S}(t) \\ \bar{S}_e(t) \end{bmatrix} - \begin{bmatrix} D_g^* & 0 & 0 \\ 0 & D_t^* & 0 \\ 0 & 0 & D_e^* \end{bmatrix} \begin{bmatrix} \bar{R}_g(t) \\ \bar{R}_t(t) \\ \bar{R}_e(t) \end{bmatrix} \qquad (15b)$$

Notice that the output variables of interest are the same as the state variables in the foregoing system models.

6. EMPIRICAL IMPLEMENTATION

The parameterization of component models, in the specific context of the sickbay at the I.I.T.Delhi was carried out,at a level of disaggregation corresponding to 13 population sub-groups and four patient health states, from the data collected from the unpublished records of the sickbay supported by information obtained through questionnaire interviews with various expert medical personnel and users of health services.

Besides aforesaid data, the calibration of the system state-model entailed reference to the log-sheets of the sickbay which furnished the behavioural pattern of the patient states from 5th Jan.1975 to 19th Jan.1975. While limitations of space prevents us from presenting full details of numerical identification of component parameters from a basis of the cross-sectional data of the base academic year 1974-75, these can be obtained from the authors. As a representative sample of empirically identified parameters we show the overall through-variable system state transition matrix of order 12x12 in table II.

As a result of the numerical solution of the model, the patient flow state trajectories during the calibration period are depicted in graph 1. The calibrated model has been used to predict the patient flows during malaria spread period from 15th June to 29th June 1975, as shown in graph 2(a) and 2(b), and during population change period from 6th to 20th July 1975, as presented in graph 3. Pertinent details regarding the 'initial patient states', input resource changes and sub-group population changes have been recorded on relevant graphs. The relevant sub-categories of IIT Population for the academic year 1974-75 are compiled in table I.

As regards the across-variable system state-model, the per unit values associated with four health states corresponding to the three dynamic components, were estimated through the celebrated Delphi technique. The empirical implmenetation of this model for predictive use is currently in progress at our Institute and will be reported in sequel to this paper.

7. CONCLUDING REMARKS

A socio-economic system theoretic approach, which offers rigorous and consistent frame-work for systems model formulation has been used successfully in this paper for developing a theoretical but empirically implementable health care system state-model Because of its firm foundation in linear graph theory, this approach is particularly amenable to automatic formulation and simulation on a digital computer.

The computational burden in this system formulation is considerably eased in so far that the state model developed is in relatively decoupled format for the through and across state-variables enabling analytical or simulation studies for these sets of variables to be accomplished separately in sequence from a knowledge of the relevant (through or across) initial patient states and the choice of the time functions for appropriate exogenous input resources (control variables) including various subgroup population projections. The patient-state concept and the transition parameters introduced in this model provide ready-measures for judging the quality of treatment service and/or the effective utility of resources in the health system.

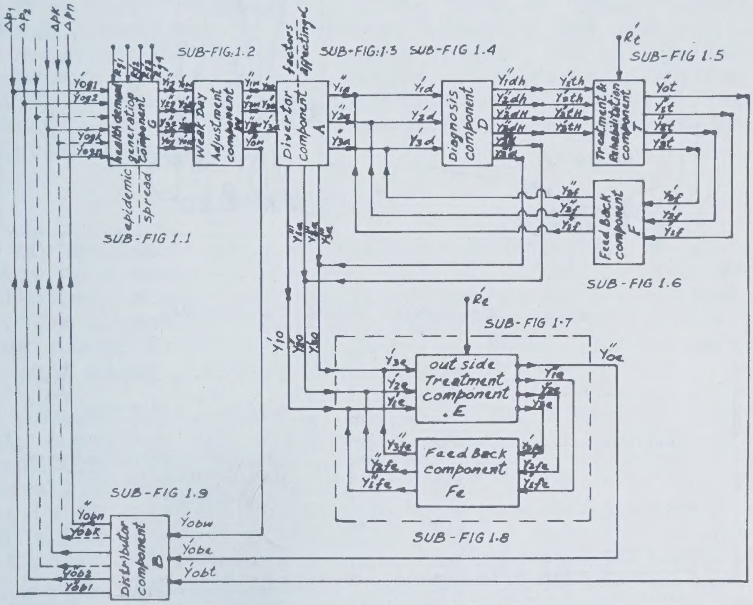
The computed solutions for the flow of patients corresponding to three periods of study show good agreement with the observed values. It has been the overriding aim throughout the development to come up with a suitable model for day-to-day use, amenable to empirical implementation with the currently available statistics. Otherwise, the framework of analysis could have been made vastly more flexible, general and powerful, if so desired. The paper opens the door for a series of simulation studies of feasible options and impacts towards an optimal design of our health care system.

REFERENCES

[1] F.D.Kennedy,"The development of a community health service system simulation model",IEEE system Science & Cybernetics Conf.,1968,California.

[2] P.K.Baokar and P.S.Satsangi,"Network Theoretic Modelling of a Health Care Delivery System",IEEE SMC Conf.1974,Dallas, Texas,Oct.1974.

[3] H.E.Koenig,et.al.,"Analysis of discrete physical systems", McGraw-Hill,New-York,1961.



ΔP1
ΔP2
ΔPK
ΔPn
SUB-FIG 1.1
health demand generation component G
epidemic spread
SUB-FIG 1.2
Week Day Adjustment component
SUB-FIG 1.3
Divertor component A
factors affecting
SUB-FIG 1.4
Diagnosis component D
SUB-FIG 1.5
Treatment & Rehabilitation component T
R'c
SUB-FIG 1.6
Feed Back component F
SUB-FIG 1.7
R'e
out side Treatment component E
Feed Back component Fe
SUB-FIG 1.8
SUB-FIG 1.9
Distributor component B
Y'obw
Y'obe
Y'obt
Y'oe
Y'ot

FIG: 1 COMMUNITY HEALTH CARE SYSTEM SCHEMATIC AND FLOW DIAGRAM.

TABLE-I : SUB-CATEGORIES OF I.I.T. POPULATION - ACADEMIC YEAR 1974-75

S No	Major Gr. I	Major Group – II				Major Group – III				Major Group – IV			
–	Age group	Age - groups				Age - groups				Age - groups			
–	16-50 yrs.	0-15yrs.	16-50 yrs		Above 50 yrs.	0-15yrs	16-50 yrs		Above 50 yrs	0-15 yrs	16 - 50 yrs		Above 50 yrs
			Male	Female			Male	Female			Male	Female	
–	Sub cat. I	Sub.Cat. II	Sub cat III	Sub cat IV	Sub cat. V	Sub cat VI	Sub. cat. VII	Sub cat VIII	Sub. cat. IX	Sub. cat. X	Sub cat. XI	Sub cat XII	Sub cat XIII
1.	1825	386	386	362	185	1202	763	730	514	1050	504	444	272

TABLE-II: NUMERICALLY IDENTIFIED THROUGH-VARIABLE SYSTEM STATE TRANSITION MATRIX (MONDAY THROUGH FRIDAY)

$$\left[\begin{array}{cccc|cccc|cccc}
\begin{matrix}0.98518\\0.01106\\0.00340\\0.00037\end{matrix} & 0.0 & 0.0 & 0.0 & \begin{matrix}0.97846\\0.01688\\0.00434\\0.00031\end{matrix} & 0.0 & 0.0 & 0.0 & \begin{matrix}0.97653\\0.02012\\0.00311\\0.00024\end{matrix} & 0.0 & 0.0 & 0.0 \\
\hline
0.0 & \multicolumn{3}{c|}{\begin{matrix}0.66444 & 0.18968 & 0.00018\\0.31458 & 0.39525 & 0.04357\\0.00098 & 0.26339 & 0.04375\\0.00000 & 0.02972 & 0.08750\end{matrix}} & 0.0 & \multicolumn{3}{c|}{\begin{matrix}0.65850 & 0.20753 & 0.00025\\0.33850 & 0.43942 & 0.06225\\0.00225 & 0.30206 & 0.06250\\0.00075 & 0.03392 & 0.12500\end{matrix}} & 0.0 & 0.0 & 0.0 & 0.0 \\
\hline
0.0 & \multicolumn{3}{c|}{\begin{matrix}0.01376 & 0.04830 & 0.25905\\0.00622 & 0.04647 & 0.20955\\0.00002 & 0.02317 & 0.20799\\0.00000 & 0.00402 & 0.14850\end{matrix}} & 0.0 & 0.0 & \multicolumn{2}{c|}{\begin{matrix}0.00676 & 0.23550\\0.00651 & 0.19050\\0.00325 & 0.18900\\0.00056 & 0.13500\end{matrix}} & 0.0 & \multicolumn{3}{c}{\begin{matrix}0.68275 & 0.39363 & 0.31190\\0.31475 & 0.38179 & 0.25470\\0.00175 & 0.19079 & 0.25270\\0.00075 & 0.03379 & 0.18070\end{matrix}}
\end{array}\right]$$

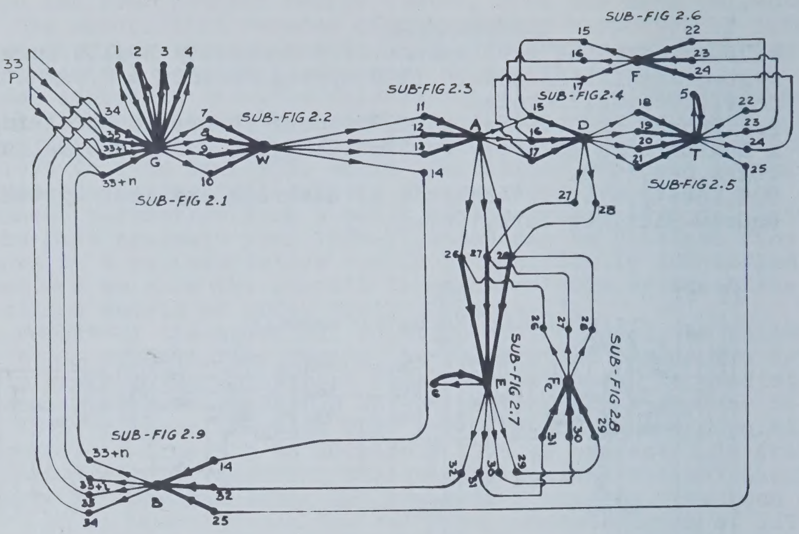
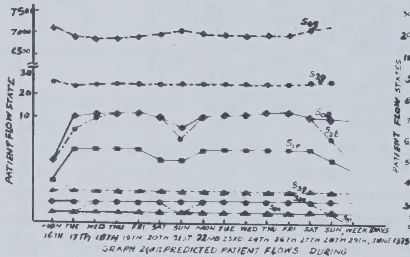
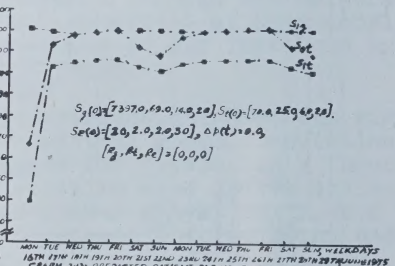

SUB-FIG 2.1
SUB-FIG 2.2
SUB-FIG 2.3
SUB-FIG 2.4
SUB-FIG 2.5
SUB-FIG 2.6
SUB-FIG 2.7
SUB-FIG 2.8
SUB-FIG 2.9
33 P
1 2 3 4
34
35
33+1
33+n
G
7 8 9 10
W
11 12 13 14
A
15 16 17
D
18 19 20 21
T
5
22 23 24 25
F
27 28
26 27 28
E
6
Fe
29 30 31 32
B

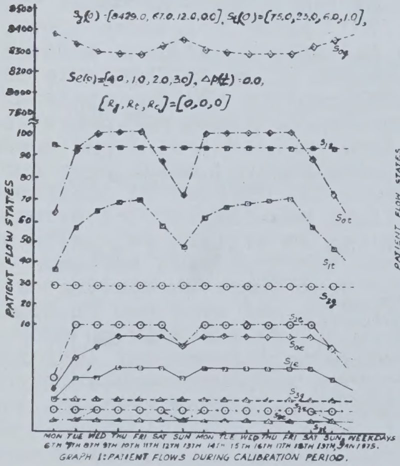
FIG: 2 COMMUNITY HEALTH CARE SYSTEM — SYSTEM GRAPH



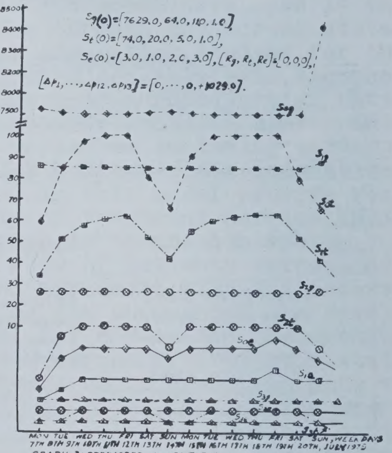
GRAPH 2(a): PREDICTED PATIENT FLOWS DURING EPIDEMIC SPREAD PERIOD



GRAPH 2(b): PREDICTED PATIENT FLOWS DURING EPIDEMIC SPREAD PERIOD



GRAPH 1: PATIENT FLOWS DURING CALIBRATION PERIOD.



GRAPH 3: PREDICTED PATIENT FLOWS DURING POPULATION CHANGE PERIOD.

ENVIRONMENTAL - ECONOMIC SYSTEMS MODELLING

D.K.BHARGAVA[1], P.S.SATSANGI and J.B.ELLIS
Electrical Engineering Department & School of Systems Studies, Indian Institute of Technology, Delhi
New Delhi-110029, India
Faculty of Environmental Studies, York University, Downsview, Ontario (Canada)

ABSTRACT

A general state-space model for an environmental-economic system from a multidimensional stand point of dynamic, spatial structural dimensions is developed following the component-to-system construct of physical system theory. The systems modelling framework is particularly simple although it incorporates such worthy features as inventory concept, nonlinear economies (for example, of scale) and internalization and externalization of antipollution costs.

1. INTRODUCTION

One of the many issues attracting the attention of the world community as a whole, during the last few years, is environmental pollution- a negative benefit accompanying the economic progress. The presence of negative externalities will prove in the long run to be of extreme danger to human health, safety, and in some cases, a threat to human survival [1] .

The present paper develops a theoretical construct that explicitly considers the environmental pollution together with the traditional economic activity system. The development views the overall system as being divided into a number of geographic regions (with suitable interregional transactions amongst them) each of which is further subdivided into a number of economic and antipollution activity sectors (wih suitable intersectoral trsnsactions between them). Each sector is considered as a collection of interconnected multiterminal components. The pertinent behavioural characteristics of each N-terminal component are completely specified by a set of N-1 'terminal equations' and a 'terminal graph' with N-1 edges. Each edge of the terminal graph has associated with it one distinct pair of complementary variables appearing in the terminal equations: a through variable Y, the flow-rate of commodities valued in constant monetary terms (e.g. Rs, \$,£) in case of economic flows and in terms of mass (e.g. gms.) in case of pollution flows; and an across variable X, the associated unit price index. The final state model is developed by suitably combining the component terminal equations with the interconnection constraints (resulting from component interconnections in the system).

2. SYSTEM COMPONENTS AND INTERCONNECTIONS

The various components, subdivided under broad catagories (with detailed terminal graphs and terminal equations summarized in Table I), are briefly described below:

2.1 Production (No.mn): represent the economic production activities together with the associated waste production in terms of the economic and pollution outputs and loctal sectoral transactions and prices.

---

[1] on leave from the Deptt. of Electronics & Telecommunication Engg. Govt. Engg. College, Jabalpur 482001.

2.2 Regional Distribution (No.mn): represent the regional supply available from the current interregional receipts.

2.3 Storage (No.mn): represent the supply in reserve which (in case of deficit) together with the various regional receipts is used to meet the sectoral demands (surplus results in material accumulation or inventory formation).

2.4 Final Consumption (No.mn): represent the sum of private and public consumption and net exports as exogenously specified flow-rates.

2.5 Final Consumption Originated Pollution (No.mn): represent the pollution originating because of the final consumption.

2.6 Untreated Pollution Discharge (No.2mn): represent the part of the pollution generated by production and final consumption that is directly discharged into the environment. A part of the production originated pollution is also directed to the internalized anti-pollution center.

2.7 Antipollution (No.m(n+1)): represent the centralized and internalized antipollution activities. The cost of internalized antipollution is reflected in the pertinent production cost.

2.8 Pollution Tolerance (No.m): represent a certain maximum environmental pollution growth rate that can be tolerated by the society (environmental standards)[2] .

2.9 Pollution Information Feedback (No.mn): represent the economic production-rate adjustments to be made in response to the difference between demand and supply.

2.10 Pollution Information Feedback (No.m): represent the adjustment of the rate of centralized antipollution activity in response to the discrepancy between the actual rate of growth of environmental pollution and the pollution tolerance.

2.11 Nonlinear Economies(No.m(2n+1)): represent, among other things, economies or diseconomies of scale in production and antipollution activities.

2.12 Transportation (No.mn(m+2n+1)): account for the costs of transportation incurred at various levels of transactions(regional and sectoral, of economic commodities and pollution) in terms of relevant flows.

2.13. Retailer's Margin (No.mn): represent the profits associated with the retail trade.

The system graph shown in Fig.1 establishes the interconnection of these components (in a two region-two sector economy) such that the following interconnection constraint equations can be formulated with reference to the maximally selected formulation tree (shown in heavy lines):

A. Fundamental Cutset Equations: representing the linearly independent flow-identities or balance equations.

$Yp = U_{\alpha\beta}\ Yr$ ..(V1) $Ydp = U_{\alpha\beta}\ Yr$ ..(V2), $Yrp = U_{\alpha\beta}\ Yr$ ..(V3)

$Yrt = Yr$ ..(V4), $Yd = Yrd$ ..(V5),

$Ysd= U_{\alpha\gamma}(Ys+Ysia)+Ysa+Yf-Yrd$ ..(V5), $Yds = Ys$ ..(V7),
$Yst=Ys$ ..(V8), $Ydsia =Ysia$ ..(V9), $Ystia= Ysia$.(V10),
$Ydsa= Ysa$ ..(V11), $Ysta=Ysa$ ..(V12), $Ydf = Yf$.(V13),
$Yfr = Yf$ ..(V14), $Yrf = Yf$ ..(V15),
$Yiap= Ywi+Ywiap+ Ydiap$ ..(V16), $Yriap=Ywi+Ywiap+Ydiap$.(V17),
$Ydwp= Ywp$ ..(V18), $Ytwp=Ywp-Ywpe-Ywi$.(V19), $Ydwf=Y\omega f$.(V20)
$Ytwf= Y\omega f-Y\omega fe$ (V21), $Yap=U_{m\alpha}(Ywp+Ywf-Ywpe-Ywi-Y\omega fe)+Y\omega ap-Yu\omega$ (V22)
$Ydap= U_{m\alpha}(Ywp+Y\omega f-Ywpe\ Ywi-Ywfe)+Ywap-Yu\omega$ ..(V23),
$Yrap= U_{m\alpha}(Ywp+Y\omega f-Ywpe-Ywi-Ywfe)+ Ywap-Yu\omega$ ..(V24),
$Yde= U_{m\alpha}(Ywpe+ Ywfe)$ .. (V25), $Ydw = Yw$ ..(V26).

B. Fundamental Circuit Equations: representing various linearly independent price-cost equality and balance relationships in the system.

$Xr = -U_{\alpha\beta}{}^{T}(Xp+Xdp+Xrap)-Xrt$ ..(L1), $Xrd=Xsd-Xrd$ ..(L2),
$Xs = -U_{\alpha\gamma}{}^{T}\ Xsd-Xds-Xst$ ..(L3), $Xsia=-U_{\alpha\gamma}{}^{T}Xsd-Xdsia-Xstia$ ..(L4),
$Xsa=-Xsd-Xdsa-Xsta$ ..(L5), $Xf=-Xsd-Xdsa-Xsta$ ..(L6),
$Xwi=U_{m\alpha}{}^{T}(Xap+Xdap+Xrap)+Xtwp-Xiap-Xriap$ ..(L7),
$Xwiap=-(Xiap+Xriap)$ ..(L8), $Xdiap=-(Xiap+Xraip)$ ..(L9),
$Xwp=-U_{m\alpha}{}^{T}(Xap+Xdap+Xrap)-Xdwp-Xtwp$ ..(L10),
$Xwpe= U_{m\alpha}{}^{T}(Xap+Xdap+Xrap-Xde)+Xtwf$ ..(L11),
$Xwf = -U_{m\alpha}{}^{T}(Xap+Xdap+Xrap)-Xdwf-Xtwf$ ..(L12),
$Xwfe= U_{m\alpha}{}^{T}(Xap+Xdap+Xrap-Xde)+Xuw$ ..(L13),
$Xwap= -Xap-Xdap-Xrap$ ..(L14), $Xuw= Xap+Xdap + Xrap$..(L15),
$Xw = -Xdw$ ..(L16).

These constraint equations together with the component terminal equations characterize the system completely. There are well established general procedures of system analysis in the literature [3 ,4] which can be used for the formulation of a state-space model for the system under consideration. Granting that the system is determinate, throughout this analytical process, one is continually assured of the existence of the matrix and functional inverses involved, due to the linear independence of the combined set of f-cutset, f-circuit and terminal equations. In the following section we present a sequence of steps which should lead to the formulation of a state space model in which the state variables constitute, a subset of the primary variables (tree-branch across variables and cotree-link through variables).

3. STATE VARIABLE FORMULATION: A GENERAL STEP-WISE PROCDURE

Step 1: Select a maximal formulation tree [5] which leads to a set of fundamental constraint equations (f-cutset and f-circuit equations)

Step 2: Write down the direct sum of dynamic terminal equations expressing derivatives of the primary variables[2] (the set S) in terms of primary variables, secondary variables and/or their derivatives and exogenous variables (inclusive of their derivatives, if any).

---

[2] In cases where a dynamic terminal equation does not involve derivatives of any primary variable it should be written explicit in an appropriate primary variable itself.

Step 3: Utilize interconnection constraint equations to eliminate secondary variables.

Step 4: Retaining primary variables (including their derivatives) in set S and exogenous variables, eliminate the remaining primary variables using the remaining set of algebraic terminal relations.

Step 5: Repeat steps 3 and 4 alternately till we are left with equations involving just the primary variables in S and exogenous variables.

Step 6: Rewrite the above set of equations (the set E) with all the derivative terms on one side. There may exist a subset S_2 of primary variables ($S_2 \subset S$) whose derivatives are not involved in this representation. In such a case it should be possible to reduce an appropriate subset E_2 of equations through elementary transformations such that it expresses the set S_2 in terms of the remaining set S_1 of primary variables (where $S_1 \subset S$ such that $S_1 \cup S_2 = S$), their derivatives and exogenous variables. The subset S_1 then constitutes the set of state variables and the subset E_1 (where $E_1 \subset E$ such that $E_1 \cup E_2 = E$) is the required set of state equations.

4. STATE MODEL FORMULATION FOR ENVIRONMENTAL -ECONOMIC SYSTEM

Further simplification accrues in our case as a result of the particular choice of component terminal characteristics and formulation tree in the sense that the dynamic equations involving through variables (T1.1, T8.1, T9.1, T11.1 and T12.1) are decoupled from (i.e. independent of) across variables and remain decoupled throughout the formulation steps 2 through 6. Note that step 1 has already been completed with reference to system graph in Fig.1 and associated constraint equations (N) and (L).

4.1 Flow-rate state model:

Step 2: Writing T1.1, T8.1, T9.1, T11.1 and T12.1 as a direct sum, we have

$$\begin{bmatrix} Y_s \\ Y_{sia} \\ Y_{sa} \\ Y_{rd} \\ Y_{u\omega} \end{bmatrix} = \begin{bmatrix} B_p & 0 & 0 & 0 & 0 \\ 0 & B_{iap} & 0 & 0 & 0 \\ 0 & 0 & B_{ap} & 0 & 0 \\ 0 & 0 & 0 & -G_{p1} & 0 \\ 0 & 0 & 0 & 0 & G_{ap1} \end{bmatrix} \begin{bmatrix} \dot{Y}_p \\ \dot{Y}_{iap} \\ \dot{Y}_{ap} \\ \dot{Y}_{dp} \\ \dot{Y}_{dap} \end{bmatrix} + \begin{bmatrix} A_p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & A_{iap} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{ap} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -G_{p2} & 0 & U_{\alpha\gamma} & U_{\alpha\gamma} & U_{\alpha\alpha} & 0 \\ 0 & 0 & 0 & 0 & G_{ap1} & 0 & 0 & 0 & -U_{mm} \end{bmatrix} \begin{bmatrix} Y_p \\ Y_{iap} \\ Y_{ap} \\ Y_{dp} \\ Y_{dap} \\ Y_{ds} \\ Y_{dsia} \\ Y_{dsa} \\ Y_{de} \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ U_{\alpha\alpha} & 0 \\ 0 & U_{mm} \end{bmatrix} \begin{bmatrix} Y_{df} \\ Y_{d\omega} \end{bmatrix}$$

where. $\alpha = mn$, $\beta = m^2 n$, $\gamma = mn^2$

$U_{\alpha\alpha}$ stands for an identity matrix of order α

U_{mm} stands for an identity matrix of order m

$U_{\alpha\beta} = [U_{\alpha\alpha} \; U_{\alpha\alpha} \text{ -m times } U_{\alpha\alpha}]^T$

$U_{\alpha\gamma} = [U_{\alpha\alpha} \; U_{\alpha\alpha} \text{ -n times } U_{\alpha\alpha}]^T$

Step 3: Using V1, V2, V7, V9, V11, V13, V16, V18, V20, V22, V25 and V26 in conjunction with T1·2, T1·3, T5·1 T6·1, T6·2, T7·1, T8·1 and T9·2 and rearranging, we have

$$\begin{bmatrix} U_{\alpha\gamma} & 0 & 0 & 0 & 0 \\ 0 & U_{\alpha\gamma} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & A_{ap}E_{ap} \\ -U_{\alpha\gamma} & -U_{\alpha\gamma} & -U_{\alpha\alpha} & U_{\alpha\alpha} & 0 \\ 0 & 0 & 0 & 0 & U_{mm}+G_{ap2}E_{ap} \end{bmatrix} \begin{bmatrix} Y_s \\ Y_{sia} \\ Y_{sa} \\ Y_{rd} \\ Y_{u\omega} \end{bmatrix} = \begin{bmatrix} B_p U_{\alpha\beta} & 0 \\ B_{iap} C_{iap} U_{\alpha\beta} & 0 \\ B_{ap} C_{ap} U_{\alpha\beta} & -B_{ap}E_{ap} \\ -G_{p1}U_{\alpha\beta} & 0 \\ G_{ap1} C_{ap} U_{\alpha\beta} & G_{ap1}E_{ap} \end{bmatrix} \begin{bmatrix} \dot Y_r \\ \dot Y_{u\omega} \end{bmatrix}$$

$$+ \begin{bmatrix} A_p U_{\alpha\beta} \\ A_{iap} C_{iap} U_{\alpha\beta} \\ A_{ap} C_{ap} U_{\alpha\beta} \\ -G_{p2} U_{\alpha\beta} \\ (G_{ap2} C_{ap} - C_{de}) U_{\alpha\beta} \end{bmatrix} Y_r + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ A_{ap}D_{ap} & B_{ap}D_{ap} & 0 \\ 0 & 0 & 0 \\ G_{ap2}D_{ap}-D_{de} & G_{ap1}D_{ap} & U_{mm} \end{bmatrix} \begin{bmatrix} Y_f \\ \dot Y_f \\ Y_{\omega} \end{bmatrix}$$

where, $C_{ap} \triangleq (U_{mm} - A_{\omega ap})^{-1} U_{m\alpha} (U_{\alpha\alpha} - A_{\omega pe} - A_{\omega i}) A_{\omega p}$, $U_n = [1\,1 \underbrace{\cdots}_{n\text{ times}} 1]$

$U_{m\alpha} = \mathrm{diag}(U_n, U_n, \underbrace{\cdots}_{m\text{ times}}, U_n)$

$D_{ap} \triangleq (U_{mm} - A_{\omega ap})^{-1} U_{m\alpha} (U_{\alpha\alpha} - A_{\omega fe}) A_{\omega f}$, $C_{iap} = (U_{\alpha\alpha} - A_{\omega iap})^{-1} A_{\omega i} A_{\omega p}$

$E_{ap} \triangleq (U_{mm} - A_{\omega ap})^{-1}$, $C_{de} \triangleq U_{m\alpha} A_{\omega pe} A_{\omega p}$, $D_{de} \triangleq U_{m\alpha} A_{\omega fe} A_{\omega f}$

Step 4: Using T2·1, T4 and T10 and rearranging, we have

$$\begin{bmatrix} U_{\alpha\gamma} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & U_{\alpha\alpha} & 0 & A_{ap}E_{ap} \\ -U_{\alpha\gamma} & -U_{\alpha\gamma} & -U_{\alpha\alpha} & U_{\alpha\alpha} & 0 \\ 0 & 0 & 0 & 0 & U_{mm}+G_{ap2}E_{ap} \end{bmatrix} \begin{bmatrix} Y_s \\ Y_{sia} \\ Y_{sa} \\ Y_{rd} \\ Y_{u\omega} \end{bmatrix} = \begin{bmatrix} B_p U_{\alpha\beta} A_d & 0 \\ B_{iap} C_{iap} U_{\alpha\beta} A_d & 0 \\ B_{ap} C_{ap} U_{\alpha\beta} A_d & -B_{ap}E_{ap} \\ -G_{p1}U_{\alpha\beta} A_d & 0 \\ G_{ap1} C_{ap} U_{\alpha\beta} A_d & G_{ap1}E_{ap} \end{bmatrix} \begin{bmatrix} \dot Y_d \\ \dot Y_{u\omega} \end{bmatrix}$$

$$\begin{bmatrix} A_p U_{\alpha\beta} A_d + B_p U_{\alpha\beta} \dot A_d \\ A_{iap} C_{iap} U_{\alpha\beta} A_d + B_{iap} C_{iap} U_{\alpha\beta} \dot A_d \\ A_{ap} C_{ap} U_{\alpha\beta} A_d + B_{ap} C_{ap} U_{\alpha\beta} \dot A_d \\ -G_{p2} U_{\alpha\beta} A_d - G_{p1} U_{\alpha\beta} \dot A_d \\ (G_{ap2} C_{ap} - C_{de}) U_{\alpha\beta} A_d + G_{ap1} C_{ap} U_{\alpha\beta} \dot A_d \end{bmatrix} Y_d + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ A_{ap}D_{ap} & B_{ap}D_{ap} & 0 \\ 0 & 0 & 0 \\ G_{ap2}D_{ap}-D_{de} & G_{ap1}D_{ap} & U_{mm} \end{bmatrix} \begin{vmatrix} F_f(t) \\ \dot F_f(t) \\ F_{\omega}(t) \end{vmatrix}$$

Step 5: Using V6, we obtain

$$\begin{bmatrix} U_{\alpha\gamma} & 0 & 0 & -A_p U_{\alpha\beta} A_d - B_p U_{\alpha\beta} \dot A_d & 0 \\ 0 & U_{\alpha\gamma} & 0 & -A_{iap} C_{iap} U_{\alpha\beta} A_d - B_{iap} C_{iap} U_{\alpha\beta} \dot A_d & 0 \\ 0 & 0 & U_{\alpha\alpha} & -A_{ap} C_{ap} U_{\alpha\beta} A_d - B_{ap} C_{ap} U_{\alpha\beta} \dot A_d & A_{ap}E_{ap} \\ -U_{\alpha\gamma} & -U_{\alpha\gamma} & -U_{\alpha\alpha} & U_{\alpha\alpha} + G_{p2} U_{\alpha\beta} A_d + G_{p1} U_{\alpha\beta} \dot A_d & 0 \\ 0 & 0 & 0 & -(G_{ap2} C_{ap} - C_{de}) U_{\alpha\beta} A_d - G_{ap1} C_{ap} U_{\alpha\beta} \dot A_d & U_{mm}+G_{ap2}E_{ap} \end{bmatrix} \begin{bmatrix} Y_s \\ Y_{sia} \\ Y_{sa} \\ Y_{rd} \\ Y_{u\omega} \end{bmatrix} =$$

$$\begin{bmatrix} B_p U_{\alpha\beta} A_d & 0 \\ B_{iap} C_{iap} U_{\alpha\beta} A_d & 0 \\ B_{ap} C_{ap} U_{\alpha\beta} A_d & -B_{ap}E_{ap} \\ -G_{p1}U_{\alpha\beta} A_d & 0 \\ G_{ap1} C_{ap} U_{\alpha\beta} A_d & G_{ap1}E_{ap} \end{bmatrix} \begin{bmatrix} \dot Y_{rd} \\ \dot Y_{u\omega} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ A_{ap}D_{ap} & 0 & 0 \\ 0 & 0 & 0 \\ G_{ap2}D_{ap}-D_{de} & G_{ap1}D_{ap} & U_{mm} \end{bmatrix} \begin{bmatrix} F_f(t) \\ \dot F_f(t) \\ F_{\omega}(t) \end{bmatrix}$$

Step 6: Transferring the derivative terms on LHS, collecting all other terms on RHS and premultiplying by $\begin{bmatrix} U_{\alpha\gamma} & U_{\alpha\gamma} & U_{\alpha\alpha} & U_{\alpha\alpha} & 0 \\ 0 & 0 & 0 & 0 & U_{mm} \end{bmatrix}$, we get

$$\begin{bmatrix} (G_{p1} - U_{\alpha\beta} B_p - U_{\alpha\gamma} B_{iap} C_{iap} - B_{ap} C_{ap}) U_{\alpha\beta} A_d & B_{ap} E_{ap} \\ G_{ap1} C_{ap} U_{\alpha\beta} A_d & G_{ap1} E_{ap} \end{bmatrix} \begin{bmatrix} \dot{Y}_{rd} \\ \dot{Y}_{uw} \end{bmatrix} =$$

$$- \begin{bmatrix} U_{\alpha\alpha} + (G_{p2} - U_{\alpha\gamma} A_p - U_{\alpha\gamma} A_{iap} C_{iap} - A_{ap} C_{ap}) U_{\alpha\beta} A_d \\ \quad + (G_{p1} - U_{\alpha\beta} B_p - U_{\alpha\gamma} B_{iap} C_{iap} - B_{ap} C_{ap}) U_{\alpha\beta} \dot{A}_d & - A_{ap} E_{ap} \\ (G_{ap2} C_{ap} - C_{de}) U_{\alpha\beta} A_d + G_{ap1} C_{ap} U_{\alpha\beta} \dot{A}_d & -(U_{mm} + G_{ap2} E_{ap}) \end{bmatrix} \begin{bmatrix} Y_{rd} \\ Y_{uw} \end{bmatrix}$$

$$+ \begin{bmatrix} A_{ap} D_{ap} & B_{ap} D_{ap} & 0 \\ D_{de} - G_{ap2} D_{ap} & - G_{ap1} D_{ap} & - U_{mm} \end{bmatrix} \begin{bmatrix} F_f(t) \\ \dot{F}_f(t) \\ F_w(t) \end{bmatrix}$$

or, $P \begin{bmatrix} \dot{Y}_{rd} \\ \dot{Y}_{uw} \end{bmatrix} = - Q \begin{bmatrix} Y_{rd} \\ Y_{uw} \end{bmatrix} + N \begin{bmatrix} F_f(t) \\ \dot{F}_f(t) \\ F_w(t) \end{bmatrix}$ where, P, Q and N have obvious interpretations.

Assuming that P is nonsingular, we finally obtain the required flow-rate state equations as

$$\begin{bmatrix} \dot{y}_{rd} \\ \dot{y}_{uw} \end{bmatrix} = - A \begin{bmatrix} y_{rw} \\ y_{uw} \end{bmatrix} + B \begin{bmatrix} F_f(t) \\ \dot{F}_f(t) \\ F_w(t) \end{bmatrix} \triangleq \begin{bmatrix} L_{rd}(y_{rd}, y_{uw}, F_f(t), \dot{F}_f(t), F_w(t)) \\ L_{uw}(y_{rd}, y_{uw}, F_f(t), F_f(t), F_w(t)) \end{bmatrix} \quad \text{(SI)}$$

where, $A = P^{-1} Q$, $B = P^{-1} N$ and L_{rd} & L_{uw} are linear functions of their arguments that can be easily defined in terms of the elements of A and B. Additionally, by appropriate substitution of the foregoing state equations into the first 3 equations at step 5, the output equations for the flow-rate state model can be written as

$$\begin{bmatrix} Y_s \\ Y_{sia} \\ Y_{sa} \end{bmatrix} = \begin{bmatrix} L_s(Y_{rd}, Y_{uw}, F_f(t), \dot{F}_f(t), F_w(t)) \\ L_{sia}(Y_{rd}, Y_{uw}, F_f(t), \dot{F}_f(t), F_w(t)) \\ L_{sa}(Y_{rd}, Y_{uw}, F_f(t), \dot{F}_f(t), F_w(t)) \end{bmatrix} \quad \text{(OI)}$$

where, L_s, L_{sia} and L_{sa} are linear functions of their arguments that can be easily defined in terms of the matrices used above.

4·2 Price State Model : In an analogous manner to the flow-rate state model developed above, we proceed through,

Step 2. (Using T1·4, T3, T8·3 and T9·3)

Step 3. (Using V5, L3, L4 and L5)

Step 4. (Using T4, T8·3, T9·3, T11·2, T11·3, T11·4, T17, T18, T19 and OI)

Step 5. (Using V8, V10, V16, V17 and V12)

Step 6. to obtain

$$\begin{bmatrix} U_{\alpha\alpha} \\ -B_p^T U_{\alpha\gamma}^T \\ -B_{iap}^T U_{\alpha\gamma}^T \\ -B_{ap}^T U_{\alpha\gamma}^T \end{bmatrix} \dot{X}_{sd} = \begin{bmatrix} H_{sd} \\ A^T U_{\alpha\gamma}^T X_{sd} + A_p^T H_{st} + B_p^T K_{st} + X_{iap} + H_{riap} - X_p \\ A_{iap}^T U_{\alpha\gamma}^T X_{sd} + A_{iap}^T H_{stia} + B_{iap}^T K_{stia} - X_{iap} \\ A_{ap}^T X_{sd} + A_{ap}^T H_{sta} + B_{ap}^T K_{sta} - X_{ap} \end{bmatrix} \quad \text{(P)}$$

where, $F_{sd}(X_{sd}, U_{\alpha\gamma}(L_s + L_{sia}) + L_{sa} + F_f(t) - y_{rd}) \triangleq H_{sd}(X_{sd}, y_{rd}, y_{uw}, F_f(t), \dot{F}_f(t), F_w(t))$

$$F_{st}(y_s) \triangleq H_{st}(Y_{rd}, Y_{u\omega}, F_f(t), \dot{F}_f(t), F_w(t))$$

$$\dot{F}_{st}(y_s) \triangleq K_{st}(Y_{rd}, Y_{u\omega}, F_f(t), \dot{F}_f(t), \ddot{F}_f(t), F_w(t), \dot{F}_w(t))$$

$$F_{stia}(y_{sia}) \triangleq H_{stia}(Y_{rd}, Y_{u\omega}, F_f(t), \dot{F}_f(t), F_w(t))$$

$$\dot{F}_{stia}(y_{sia}) \triangleq K_{stia}(Y_{rd}, Y_{u\omega}, F_f(t), \dot{F}_f(t), \ddot{F}_f(t), F_\omega(t), \dot{F}_\omega(t))$$

$$F_{sta}(Y_{sa}) \triangleq H_{sta}(Y_{rd}, Y_{u\omega}, F_f(t), \dot{F}_f(t), F_w(t))$$

$$\dot{F}_{sta}(Y_{sa}) \triangleq K_{sta}(Y_{rd}, Y_{u\omega}, F_f(t), \dot{F}_f(t), \ddot{F}_f(t), F_w(t), \dot{F}_w(t))$$

$$C_{riap} \triangleq C_{iap} U_{\alpha\beta} A_d \quad \text{and} \quad F_{riap}(C_{iap} y_{rd}) \triangleq H_{riap}(Y_{rd})$$

Premultiplication by $[\,U_{\alpha\alpha} \;\; 0 \;\; 0 \;\; 0\,]$ then immediately gives

$$\dot{X}_{sd} = H_{sd} \qquad (S\text{II})$$

which constitutes the price state equation. The bottom three equations from (P) can then be readily written in the form of the output equation for the price state model as:

$$\begin{bmatrix} X_p \\ X_{iap} \\ X_{ap} \end{bmatrix} = \begin{bmatrix} H_p(Y_{rd}, Y_{u\omega}, F_f(t), \dot{F}_f(t), \ddot{F}_f(t), F_w(t), \dot{F}_w(t)) \\ H_{iap}(Y_{rd}, Y_{u\omega}, F_f(t), \dot{F}_f(t), \ddot{F}_f(t), F_w(t), \dot{F}_w(t)) \\ H_{ap}(Y_{rd}, Y_{u\omega}, F_f(t), \dot{F}_f(t), \ddot{F}_f(t), F_w(t), \dot{F}_w(t)) \end{bmatrix} \qquad (O\text{II})$$

5. CONLUDING REMARKS

The complete state model consists of state equations (SI) and (SII) and output equations (OI) and (OII) and is characterized in terms of economic state variables Yrd and Xsd representing respectively material flow-rate availabilities and par unit prices of materials stored in various activity sectors of different geographic regions together with the pollution state variable Yuw representing the uneliminated pollution discharges from the centralized antipollution industry in different geographic regions. Final consumption and pollution tolerance are exogenous. Obviously all other endogenous economic variables of interest (such as Yp, denoting gross outputs of various production sectors in different geographic regions) can be readily computed in terms of the state and exogenous variables.

It is worth noting that the high degree of sophistication built into the general environmental-economic system model presented in this paper is due primarily to (i) suitable relaxation of assumptions of idealized industrial firm type of component characterizations [4] ;(ii) inclusion of technological economies of scale in production, transportation and antipollution activities as an additive nonlinearity in cost equations [6] ; (iii) explicit incorporation, on the economic front, of both the economic aspects of supply and demand in a feedback framework that allows adjustment of sectoral growth rates in response to the difference between demand and supply [7] and (iv) explicit consideration, on the environmental front, of the necessity of adjusting the growth rate of the centralized antipollution industry in response to the discrepancy between the observed rate of increase of environmental pollution and pollution tolerance (or environmental standards [2]).

In spite of the provision of the aforesaid potential avenues of significant improvements in our modelling framework, the resulting state model is a particularly simple one in so far that the flow model is independent of the price model due to an appropriate choice of component terminal characteristics and maximally selected formulation tree.

In operationalizing the state model in empirical terms the data requirements are as follows:

(a) Flow-rate Model: structural parameters relating to current and fixed capital input requirements of various economic and antipollution sectors, interregional economic trading pattern, (note that the model developed above assumes that there is no inter-regional movement of pollution) waste generation by various production and final consumption activities, internalization of the antipollution activity in different sectors and the pollution discharged directly to the environment together with base year values of the state variables and time-series data on the exogenous variables.

(b) Price State Model: in addition to the data used and time profiles of states obtained in (a) we need the time series data on inventary changes and per unit cost of stored materials in order to specify the terminal characteristics of the various storage components.

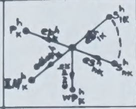
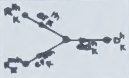
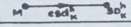
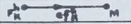
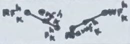
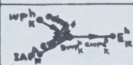
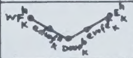
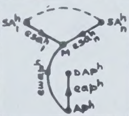
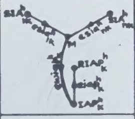
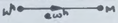
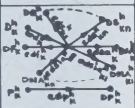
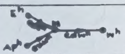
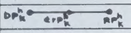
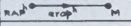
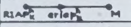
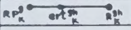
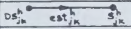
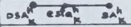
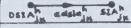
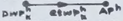
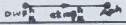
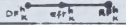
All other non-state variables can be obtained from the information available from (a) and (b) and the knowledge of the terminal characteristics of the algebraic components.

At an empirical pilot level of study the above model under implementation in terms of a 2 region-2 sector disaggregation for India.

REFERENCES

1. Ahmed, S.B., "Input-Output Analysis in Environmental Modelling" Editorial, IEEE Trans. Syst. Man Cybern., vol. SMC-3, pp. 537-538, Nov. 1973.
2. Chen. Kan, "Input-output Economic Analysis of Environmental Impact", IEEE Trans. Syst. Man Cybern., vol. SMC-3, pp. 539-547, Nov. 1973.
3. Koenig, H.E., Tokad, Y., Kesavan, H.K., and Hedges, H.G., Analysis of Discrete Physical Systems, New York: McGraw Hill, 1967.
4. Satsangi, P.S., and Ellis, J.B., "A generalized system-theoretic framework for modelling large scale national economic system in dynamic, structural and spatial terms", Int. J. Syst. Sci., Vol. 2, No. 2, pp. 213-223, 1971.
5. Satsangi, P.S., "A Physical System Theory Modelling Framework Generalized for Large Scale Economic Systems", Ph.D. Thesis, University of Waterloo, 1969.
6. Tummala, R.L., and Connor, L.J., "Mass-Energy Based Economic Models", IEEE Trans. Syst., Man and Cybern., vol. SMC-3, pp. 548-555, Nov. 1973.
7. Vogt, W.G., Mickle, M.H., Aldermenshian, H., "A Dynamic Leontief Model for a Productive System", Proc. IEEE, Vol. 63, No. 3, pp. 438-443, March 1975.

TABLE I SYSTEM COMPONENTS, TERMINALGRAPHS AND TERMINAL EQUATIONS

COMPONENT	TERMINALGRAPH	TERMINAL EQUATIONS	
		TYPICAL ELEMENTS	MATRIX REPRESENTATION
PRODUCTION		$Ys^h_{jK} = Ap^h_{1K} Yp^h_{1K} + Bp^h_{1K} \dot{Y}p^h_{1K}$ $Y\omega p^h_K = A\omega p^h_K Yp^h_K , \; Ydiap^h_K = 0$ $Xp^h_K = -\sum_{j=1}^{n} (Ap^h_{jK} Xs^h_{jK} + Bp^h_{jK} \dot{X}s^h_{jK}) - Xdiap^h_K$	$Ys = Ap\,Yp + Bp\,\dot{Y}p$ (T1.1) $Y\omega p = AWp\,Yp$ (T1.2) $Ydiap = 0$ (T1.3) $Xp = Ap^T Xs - Bp^T \dot{X}s - Xdiap$ (T1.4)
REGIONAL DISTRIBUTION		$Yr^{gh}_K = Ad^{gh}_K \; Yd^h_K$ $Xd^h_K = -\sum_{g=1}^{m} Ad^{gh}_K Xr^{gh}_K$	$Yr = Ad\,Yd$ (T2.1) $Xd = -Ad^T Xd$ (T2.2)
STORAGE		$\dot{X}sd^h_K = Fsd^h_K(Xsd^h_K, Ysd^h_K)$	$\dot{X}sd = Fsd(Xsd, Ysd)$ (T3)
FINAL CONSUMPTION		$Yf^h_K = Ff^h_K(t)$	$Yf = Ff(t)$ (T4)
FINAL CONSUMPTION ORIGINATED POLLUTION		$Y\omega f^h_K = A\omega f^h_K \; Yrf^h_K$ $Xrf^h_K = 0$	$Y\omega f = A\omega f\,Yrf$ (T5.1) $Xrf = 0$ (T5.2)
PRODUCTION ORIGINATED UNTREATED-CUM-INTERNALLY TREATED POLLUTION		$Y\omega pe^h_K = A\omega pe^h_K \; Yd\omega p^h_K$ $Y\omega i^h_K = A\omega i^h_K \; Yd\omega p^h_K$ $Xd\omega p^h_K = 0$	$Y\omega pe = A\omega pe\,Yd\omega p$ (T6.1) $Y\omega i = A\omega i\,Yd\omega p$ (T6.2) $Xd\omega p = 0$ (T6.3)
FINAL CONSUMPTION ORIGINATED POLLUTION DISCHARGE		$Y\omega fe^h_K = A\omega fe^h_K \; Y\omega f^h_K$ $Xd\omega f^h_K = 0$	$Y\omega fe = A\omega fe\,Yd\omega f$ (T7.1) $Xd\omega f = 0$ (T7.2)
CENTRALIZED ANTIPOLLUTION (CAP)		$Ysa^h_K = Aap^h_K Yap^h + Bap^h_K \dot{Y}ap^h$ $Y\omega ap^h = A\omega ap^h Yap^h$ $Xap^h = -\sum_{K=1}^{n} (Aap^h_K Xsa^h_K + Bap^h_K \dot{X}sa^h_K)$	$Ysa = Aap\,Yap + Bap\,\dot{Y}ap$ (T8.1) $Y\omega ap = A\omega ap\,Yap$ (T8.2) $Xap = -Aap^T Xsa - Bap^T \dot{X}sa$ (T8.3)
INTERNALIZED ANTIPOLLUTION (SAP)		$Ysia^h_{jK} = Aiap^h_{jK} Yiap^h_K + Biap^h_{jK} \dot{Y}iap^h_K$ $Y\omega iap^h_K = A\omega iap^h_K \; Yiap^h_K$ $Xiap^h_K = -\sum_{j=1}^{n} (Aiap^h_{jK} Xsia^h_{jK} + Biap^h_{jK} Xsia^h_{jK})$	$Ysia = Aiap\,Yiap + Biap\,\dot{Y}iap$ (T9.1) $Y\omega iap = A\omega iap\,Yiap$ (T9.2) $Xiap = -Aiap^T Xsia - Biap^T \dot{X}sia$ (T9.3)
POLLUTION TOLERANCE		$Y\omega^h = F\omega^h(t)$	$Y\omega = F\omega(t)$ (T10)
ECONOMIC INFORMATION FEEDBACK		$Yrd^h_K = \sum_{j=1}^{n} (Yds^h_{jK} + Ydsia^h_{jK}) + Ydsa^h_K + Ydf^h_K - Gp1\,Ydp^h_K - Gp2\,Ydp^h_K$ $Xds^h_{jK} = 0 , \; j = 1, \ldots, n$ $Xdsia^h_{jK} = 0 , \; j = 1, \ldots, n$ $Xdsa^h_K = 0$ $Xdf^h_K = 0$ $Xdp^h_K = 0$	$Yrd = U_{Kr}(Yds + Ydsia) + Ydsa + Ydf - Gp1\,\dot{Y}dp - Gp2\,Ydp$ (T11.1) $Xds = 0$ (T11.2) $Xdsia = 0$ (T11.3) $Xdsa = 0$ (T11.4) $Xdf = 0$ (T11.5) $Xdp = 0$ (T11.6)
POLLUTION INFORMATION FEEDBACK		$Y\omega^h = Gap1\,\dot{Y}dap^h + Gap2\,Ydap^h - Yde^h + Yd\omega^h$ $Xdap^h = 0$ $Xde^h = 0$ $Xd\omega^h = 0$	$Ya\omega = Gap1\,\dot{Y}dap + Gap2\,Ydap - Yde + Yd\omega$ (T12.1) $Xdap = 0$ (T12.2) $Xde = 0$ (T12.3) $Xd\omega = 0$ (T12.4)
NONLINEAR ECONOMICS			
PRODUCTION		$Xrp^h_K = Frp^h_K(Yrp^h_K)$	$Xrp = Frp(Yrp)$ (T13)
CAP		$Xrap^h = Frap^h(Yrap^h)$	$Xrap = Frap(Yrap)$ (T14)
SAP		$Xriap^h_K = Friap^h_K(Yriap^h_K)$	$Xriap = Friap(Yriap)$ (T15)
REGIONAL TRANSPORTATION		$Xrt^{gh}_K = Frt^{gh}_K(Yrt^{gh}_K)$	$Xrt = Frt(Yrt)$ (T16)
SECTORAL TRANSPORTATION			
ECONOMIC - ECONOMIC		$Xst^h_{jK} = Fst^h_{jK}(Yst^h_{jK})$	$Xst = Fst(Yst)$ (T17)
ECNOMIC - CAP		$Xsta^h_K = Fsta^h_K(Ysta^h_K)$	$Xsta = Fsta(Ysta)$ (T18)
ECONOMIC - SAP		$Xstia^h_{jK} = Fstia^h_{jK}(Ystia^h_{jK})$	$Xstia = Fstia(Ystia)$ (T19)
POLLUTION TRANSPORTATION			
PRODUCTION ORIGINATED		$Xt\omega p^h_K = Ft\omega p^h_K(Yt\omega p^h_K)$	$Xt\omega p = Ft\omega p(Yt\omega p)$ (T20)
FINAL CONSUMPTION ORIGINATED		$Xt\omega f^h_K = Ft\omega f^h_K(Yt\omega f^h_K)$	$Xt\omega f = Ft\omega f(Yt\omega f)$ (T21)
RETAILERS MARGIN		$Xfr^h_K = Ffr^h_K(Yfr^h_K)$	$Xfr = Ffr(Yfr)$ (T22)

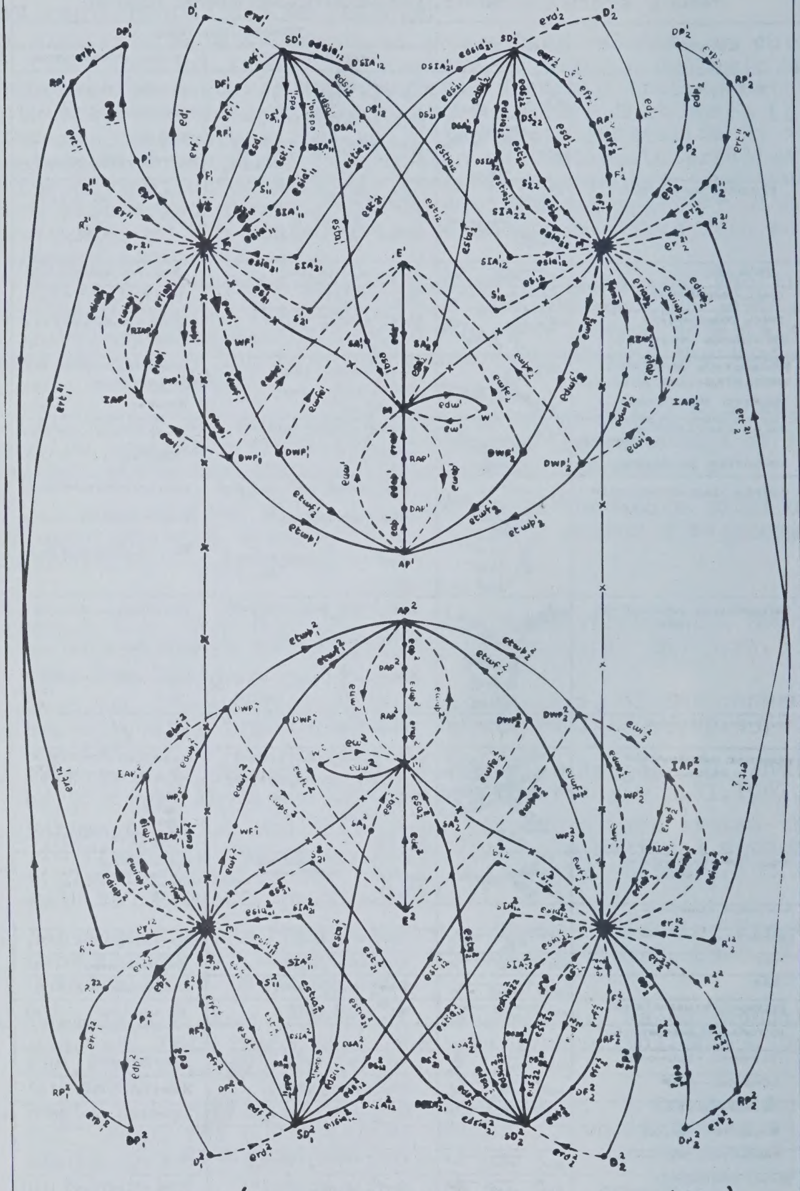

D'1
SD'1
DSIA'12
DP'1
RP'1
SD'2
D'2
DSIA'21
DP'2
RP'2
SIA'21
S'21
SIA'12
S'12
E'
W'
edw'
ew'
IAP'1
DWP'1
DWF'1
DWF'2
DWP'2
IAP'2
RAP'
DAP'
AP'
AP2
DAP2
RAP2
E2
IAP21
DWP21
DWF21
DWF22
DWP22
IAP22
RP21
DP21
SD21
D21
DSIA212
DSIA221
SD22
D22
DP22
RP22

Fig.1 The System Graph (Formulation Tree is shown by Continuous Lines)

MAXIMUM-DISTANCE QUANTIZATION FOR DETECTION

H.V. POOR AND J.B. THOMAS
Department of Electrical Engineering
Princeton University, Princeton, N.J.

ABSTRACT

The problem of optimal quantization for detection is approached using signal selection techniques. Criteria are established for the optimal selection of quantizer parameters when Ali-Silvey distances are used as measures of goodness. Results are given for both binary and local (or small-signal) detection and, for the local case, this optimum is shown to be independent of the particular Ali-Silvey distance used. Examples are presented to illustrate these concepts in both the local and general cases.

I. INTRODUCTION

When quantizers are studied as elements of detection systems, their optimization is complicated by the intractability of error probabilities. Typical quantizer design criteria such as mean-square error or entropy are unsuitable in this context since they are only weakly related to optimal detection. In [1] the performance of quantizers in detection systems is analyzed and [2] includes a brief discussion of their optimization. In [3] the problem is approached on an asymptotic, small-signal basis using Pitman efficiency and local power slope as measures of system performance.

In this paper, we circumvent problems arising from error probabilities by considering the Ali-Silvey class of distance measures. Because of their strong links to probability of error, members of this class are frequently used as criteria for signal selection. Here, we use them in a similar manner to derive tractable, if not optimal, design procedures for quantizer-detectors.

II. PRELIMINARIES

A. The Binary Detector

First we consider the binary detection problem, in which we wish to decide between two possible statistical situations on the basis of a sequence $\{x_i; i=1,\ldots,n\} \equiv \mathbf{x}$ of independent, identically distributed (i.i.d.) observations. That is to say, we have hypothesis

$$\begin{aligned} &H_0: x_i \sim P_0 \ (x_i \text{ has distribution } P_0),\ i=1,\ldots n \\ \text{versus} \quad & \\ &H_1: x_i \sim P_1,\ i=1,\ldots,n. \end{aligned} \tag{1}$$

Of course, the Neyman-Pearson lemma [4] gives the optimal detection scheme as,

$$\varphi(\underline{x}) = \begin{cases} 1, & > \tau \\ a, \sum_{i=1}^{n} \log(L(x_i)) & = \tau \\ 0, & < \tau \end{cases} \tag{2}$$

where $L(x)=dP_1(x)/dP_0(x)$, the likelihood ratio (we assume that $P_1 << P_0$); τ and a are chosen to give desired performance; and we accept hypothesis H_1 with probability $\varphi(\underline{x})$ when $\underline{x}$ is observed. Here, we wish to replace log(L) with a quantizer Q choosing the parameters of Q in some optimal way. To do so we need optimality criteria more manageable than error probabilities.

B. Ali-Silvey Distances

Notions of distances between probability measures have been widely used in the design of statistical experiments and applications of these ideas to signal selection and other communication-theoretic problems are common. A review of these concepts and results can be found in the introductory sections of [5].

An important class of distance measures which includes, among others, the J-divergence, Matsusita, and Bhattacharyya distances, was studied by Ali and Silvey in [6]. This class is defined as those distances which can be written in the form:

$$d(P_0,P_1) = f\left\{E_0\{C(L)\}\right\} \tag{3}$$

where $d(P_0,P_1)$ is the distance between the probability measures P_0 and P_1, f is an increasing function, C is a convex function on $(0,\infty)$, L is the likelihood ratio dP_1/dP_0 and E_0 denotes expectation with respect to P_0. The importance of this class arises largely from theorems closely relating maximum distance to minimum probability of error in problems such as (1) (when prior probabilities are imposed).

For reference, two important examples are given here:

$$\begin{aligned} &\text{(a) J-divergence} : J = E_0\{(L-1)\log L\} \\ &\text{(b) Matsusita's Distance} : d = [E_0\{(\sqrt{L}-1)^2\}]^{1/2} \end{aligned} \tag{4}$$

III. THE MAXIMUM-DISTANCE QUANTIZER

A. The Quantizer

An M-level quantizer Q(x) can be represented as a pair $(\underline{t},\underline{q})$, where $\underline{q} = (q_1,\ldots,q_M)$ are the levels and $\underline{t} = (t_0,\ldots,t_M)$ are the breakpoints. We require $-\infty=t_0 \le t_1 \le \ldots \le t_{M-1} \le t_M = +\infty$, and take

$$Q(x) = q_k \text{ if } x \in (t_{k-1},t_k], \quad k=1,\ldots,M. \tag{5}$$

The parameters of the quantizer are illustrated in Fig. 1.

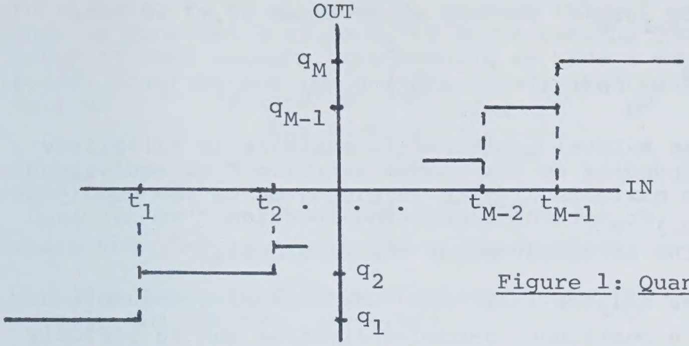


Figure 1: Quantizer Parameters

Note that, once the points $\underline{t}$ have been specified, the problem of optimal detection can be solved immediately from the Neyman-Pearson lemma. The levels $\underline{q}$ are irrelevant to optimal detection since a test may be based solely in the numbers of observations falling in each quantization interval [2]. However, certain specification of $\underline{q}$ may lead to a more practical implementation than do others. Specifically, by choosing

$$q_k = \log(P_1(t_{k-1},t_k]/P_0(t_{k-1},t_k]),\ k=1,\dots,M, \tag{6}$$

the post-quantization detector can be represented optimally (for given $\underline{t}$) in the form of Eq. (2) with log(L) replaced by Q, i.e. a threshold test based on the statistic

$$T_Q(\underline{x}) = \sum_{i=1}^{n} Q(x_i). \tag{7}$$

From now on, the symbol $\underline{q}$ will denote this particular choice of levels and is understood to be a function of $\underline{t}$.

B. Maximum-Distance Quantizers

From the above we see that the problem of optimal quantizer design for (1) is that of optimally placing the points $\underline{t}$. This could be done in a number of ways. Ideally, $\underline{t}$ should be chosen to minimize the probability of error; however, the error probability is not simply related to $\underline{t}$ for general n and M. (For fixed $\underline{t}$, the quantizer statistic (7) is multinomially distributed with known multinomial probabilities. The problems arise in that the rejection region cannot be written simply in a form that lends itself to standard optimization techniques.) The Ali-Silvey distance measures are easily expressible in terms of $\underline{t}$, and so lead to practical design criteria.

Denote by Q_0 and Q_1 the distribution of Q(x) when x has distribution P_0 and P_1, respectively. The likelihood ratio between Q_0 and Q_1 depends only on $\underline{t}$, and hence the distance in any Ali-Silvey sense depends only on $\underline{t}$ (this, of course, for fixed P_0 and P_1). Defining

$$\tilde{q}_k = \exp(q_k)\ ,\ k=1,\dots,M,$$

we can write (8)

$$L_Q(x) \overset{D}{=} dQ_1/dQ_0 = \tilde{q}_k$$

when x is in $(t_{k-1}, t_k]$. For any C convex on $(0,\infty)$ we have

$$d_C(\underline{t}) \overset{D}{=} E_{Q_0}\{C(L_Q)\} = \sum_{k=1}^{M} C(\tilde{q}_k)\; P_0(t_{k-1}, t_k] \tag{9}$$

Thus, finding an M-level quantizer to maximize an Ali-Silvey distance corresponding to the convex function C is equivalent to finding $\underline{t}$ to maximize $d_C(\underline{t})$ in (9), subject to the condition $-\infty = t_0 \le t_1 \le \ldots \le t_{M-1} \le t_M = +\infty$ (the increasing function f of (3) is irrelevant to the maximization of the distance).

To maximize $d_C(\underline{t})$ we first make the following assumptions:

(i) C has a continuous second derivative and is strictly convex in a neighborhood of unity. (An example of an Ali-Silvey distance which does not satisfy this condition is Kolmogorov's measure of variational distance, $E_0\{|L-1|\}$.)

(ii) P_0 and P_1 have continuous Lebesgue densities p_0 and p_1, respectively.

To look for maxima of $d_C(\underline{t})$, we search for points $\underline{t}$ at which the gradient of $d_C(\underline{t})$ is zero, i.e.

$$\partial d_C(\underline{t})/\partial t_k = 0 \;, \; k=1,\ldots,(M-1)$$

(note that t_0 and t_M are fixed). Under Assumptions (i) and (ii) we have

$$\partial d_C(\underline{t})/\partial t_k = p_0(t_k)\{[C(\tilde{q}_k) - \tilde{q}_k C'(\tilde{q}_k)] - [C(\tilde{q}_{k+1}) - \tilde{q}_{k+1} C'(q_{k+1})]\} + p_1(t_k)[C'(\tilde{q}_k) - C'(\tilde{q}_{k+1})] \tag{10}$$

where C' is the first derivative of C and $\tilde{q}_k$ is from (8). Setting (10) equal to zero, we have the necessary conditions for a maximum:

$$L(t_k) = \frac{p_1(t_k)}{p_0(t_k)} = \frac{(\tilde{q}_{k+1} C'(\tilde{q}_{k+1}) - C(\tilde{q}_{k+1})) - (\tilde{q}_k C'(\tilde{q}_k) - C(\tilde{q}_k))}{(C'(\tilde{q}_{k+1}) - C'(\tilde{q}_k))} \tag{11}$$

C. Example-Matsusita Distance

To illustrate Eq. (11) consider the case of Matsusita distance [Eq. (4b)], $C(L) = (\sqrt{L} - 1)^2$: Here, (11) becomes

$$L(t_k) = (\tilde{q}_k \tilde{q}_{k+1})^{1/2} \;\; ; \; k=1,\ldots(M-1)$$

or, equivalently,

$$\log(L(t_k)) = (q_k + q_{k+1})/2$$

We see that $L(t_k)$ is the geometric average of the numbers $\tilde{q}_k$ and $\tilde{q}_{k+1}$; i.e. the log-likelihood ratio at t_k is the arithmetic average of the levels q_k and q_{k+1}.

IV. THE SMALL-SIGNAL CASE

A. Problem Definition

Typically, the hypotheses H_0 and H_1 represent, respectively, the absence and presence of a signal. In this situation we are often interested in the local, or small-signal, case corresponding to allowing P_1 to approach P_0 in some natural way. To study this we consider a class $\mathcal{P}_\Theta$ of distributions indexed by an open set Θ of real numbers. As before, we have a sequence $\underline{x}$ of n i.i.d. observations and, singling out a particular $\theta_0 \epsilon \bar{\Theta}$, we wish to test

$$H_0: x_i \sim P_{\theta_0}, \; i=1,\dots,n$$

versus (12)

$$H_1: x_i \sim P_\theta, \; i=1,\dots,n$$

where $\theta \epsilon \Theta$, $\theta > \theta_0$. The local case corresponds to $\theta \to \theta_0$.

The generalized Neyman-Pearson lemma [4] leads to the locally-optimum test given by:

$$\varphi(\underline{x}) = \begin{cases} 1, & > \tau \\ a, \; \sum_{i=1}^{n} \frac{\partial}{\partial\theta} L_\theta(x_i) \Big|_{\theta=\theta_0} & = \tau \\ 0, & < \tau \end{cases} \tag{13}$$

where L_θ is the likelihood ratio* between P_θ and P_{θ_0}.

B. Local Maximum-Distance Quantizers

We want to investigate the conditions for maximum-distance quantization (11) for the problem (12) as $\theta \to \theta_0$. First we note that, for any fixed $\underline{t}$,

$$\lim_{\theta\to\theta_0} \tilde{q}_k = 1, \; k=1,\dots,M;$$

$$\lim_{\theta\to\theta_0} \frac{p_\theta(t_k)}{p_{\theta_0}(t_k)} = 1, \; k=1,\dots,(M-1);$$

and (applying L'Hopital's rule)

$$\lim_{\theta\to\theta_0} \frac{(\tilde{q}_{k+1}C'(\tilde{q}_{k+1}) - C(\tilde{q}_{k+1})) - (\tilde{q}_k C'(\tilde{q}_k) - C(\tilde{q}_k))}{(C'(\tilde{q}_{k+1}) - C'(\tilde{q}_k))} = 1, k=1,\dots,(M-1).$$

Thus, for the local case, the levels $q_k = \log \tilde{q}_k$ approach zero for all k, and the necessary conditions (11) degenerate.

To overcome this difficulty, we first note that comparing the statistic (7) to a threshold is equivalent to comparing any constant multiple to that multiple of the threshold. Thus, the optimum levels $\underline{q}$ in (6) can be scaled without changing the test, so that the Conditions (6) can be written as

$$q_{k+1}/q_k = \log(\tilde{q}_{k+1})/\log(\tilde{q}_k), \; k=1,\dots,(M-1). \tag{14}$$

---

[*] We impose the regularity conditions:
(i) $P_\theta \ll P_{\theta_0} \; \forall \theta > \theta_0$, and (ii) $\frac{\partial}{\partial\theta}\int_I L_\theta dP_{\theta_0} = \int_I \frac{\partial}{\partial\theta} L_\theta dP_{\theta_0}$, I an interval.

Taking limits, and applying L'Hopital's rule, we see that

$$\lim_{\theta\to\theta_0} q_{k+1}/q_k = \frac{\partial\tilde{q}_{k+1}}{\partial\theta} \Big/ \frac{\partial\tilde{q}_k}{\partial\theta}\Big|_{\theta=\theta_0} \tag{15}$$

The necessary condition (11) can also be written in a limiting form by the proper norming. We notice that, for a particular $\theta>\theta_0$, (11) is equivalent to:

$$\frac{L_\theta(t_k)-1}{\theta-\theta_0} = \frac{((\tilde{q}_{k+1}-1)C'(\tilde{q}_{k+1})-C(\tilde{q}_{k+1}))-((\tilde{q}_k-1)C'(\tilde{q}_k)-C(\tilde{q}_k))}{(\theta-\theta_0)\ (C'(\tilde{q}_{k+1})-C'(\tilde{q}_k))} \tag{16}$$

$$k=1,\ldots,(M-1)$$

Certainly, for a particular $\theta>\theta_0$, t satisfies (16) iff. (11) is satisfied. Taking limits as $\theta\to\theta_0$ on both sides of (16) yields a necessary relationship to be satisfied (along with (15)) by the local maximum-distance quantizer:

$$\frac{\partial}{\partial\theta} L_\theta(t_k)\Big|_{\theta=\theta_0} = \frac{1}{2}\left[\frac{\partial\tilde{q}_k}{\partial\theta} + \frac{\partial\tilde{q}_{k+1}}{\partial\theta}\right]_{\theta=\theta_0}, \quad k=1,\ldots,(M-1) \tag{17}$$

Choosing the limiting levels $\underline{q}^*$ as

$$q_k^* = \partial\tilde{q}_k/d\theta\Big|_{\theta=\theta_0}, \quad k=1,\ldots,M\ ,$$

and noting that

$$\frac{\partial\tilde{q}_k}{\partial\theta}\Big|_{\theta=\theta_0} = \int_{t_{k-1}}^{t_k} \frac{\partial L_\theta}{\partial\theta}\Big|_{\theta=\theta_0} dP_{\theta_0} \Big/ \int_{t_{k-1}}^{t_k} dP_{\theta_0}$$

we have that the local maximum-distance quantizer satisfies

$$\frac{\partial L_\theta}{\partial\theta}\Big|_{\theta=\theta_0} = (q_{k+1}^* + q_k^*)/2, \quad k=1,\ldots,(M-1) \tag{18}$$

Note that, for the local case, the maximum-distance quantizer is independent of the choice of C. This indicates that, for small signals, maximizing (9) will give similar quantizers regardless of which C is chosen.

C. Example-Known Signal in Additive Noise

If we consider the problem of local detection of a known signal in additive noise, Eqs. (18) simplify somewhat. We have a sequence $\{x_i; i=1,\ldots,n\}$ of i.i.d. samples, and want to test

$$H_0:\ x_i = h_i\ ,\quad i=1,\ldots,n$$

versus (19)

$$H_1:\ x_i = h_i + \theta\ ,\quad i=1,\ldots,n,\ \theta\to 0^+$$

where $\{h_i; i=1,\ldots,n\}$ is an i.i.d. noise sequence with a common differentiable Lebesgue density g. This fits the framework of part IV-A with $\theta_0 = 0$ and

$$L_\theta(x) = g(x-\theta)/g(x)$$

Note that
$$\left.\frac{\partial L_\theta(x)}{\partial\theta}\right|_{\theta=0} = -\,g'(x)/g(x) \overset{D}{=} g_{\ell o}(x)$$

Equation (18) becomes
$$g_{\ell o}(t_k) = (q_k^* + q_{k+1}^*)/2 \;;\; k=1,\ldots,(M-1)$$
where (20)
$$q_k^* = [g(t_{k-1}) - g(t_k)] \Big/ \int_{t_{k-1}}^{t_k} g(x)dx \;;\; k=1,\ldots,M.$$

V. EXAMPLE-GENERALIZED GAUSSIAN NOISE

A. Solutions for J-Divergence and Matsusita Distance

To illustrate the concepts developed here, we consider the known-signal/additive-noise case of (19). A convenient class of noise distributions for studies of this type is the generalized Gaussian class, defined in terms of its density functions as: $g(x|\ell) = (\ell\eta(\ell) / 2\Gamma(1/\ell))\exp\{-(\eta(\ell)|x|)^\ell\};\; x\varepsilon(-\infty,\infty)$

where $\eta(\ell) \overset{D}{=} [\Gamma(3/\ell) / \Gamma(1/\ell)]^{1/2}$,

Γ is the gamma function, and ℓ, a positive number, parameterizes the class. Each member has zero mean and unit variance, making this class a convenient one for studying the effects of density shape on detection systems. Note that $\ell = 2$ gives the standard Gaussian distribution and $\ell = 1$ gives the Laplace double-exponential distribution.

We wish to solve for the maximum-distance quantizer for the two cases of J-divergence and Matsusita distance, [see Eq. (4)]. In general, the necessary conditions (11) have no closed form solutions for these cases. However, the equations for the distance (9) and its gradient (10) can be programmed easily for conventional gradient-ascent techniques. This was done, and results for the case M=4 (i.e. 4-level quantization) are presented here.

For both the J-divergence and Matsusita distances, the max.-distance quantizer, given θ, can be characterized as follows:

$$-t_0=t_4=+\infty$$
$$t_2=\theta/2 \qquad\qquad q_4=-q_1>0 \qquad\qquad (21)$$
$$(\theta/2-t_1)=(t_3-\theta/2)>0 \qquad\qquad q_3=-q_2>0$$

The local version satisfies (21) with $\theta = 0$. The optimum quantizers are thus completely specified by the differences $(t_3-\theta/2)$ and the levels ratios q_4/q_3. Figures 2 and 3 give plots of these parameters (for various θ and for the local case) as a function of ℓ, the density parameter. Note that the results for the J-divergence and Matsusita distance are not identical, but for 'small' θ they are essentially the same, as predicted.

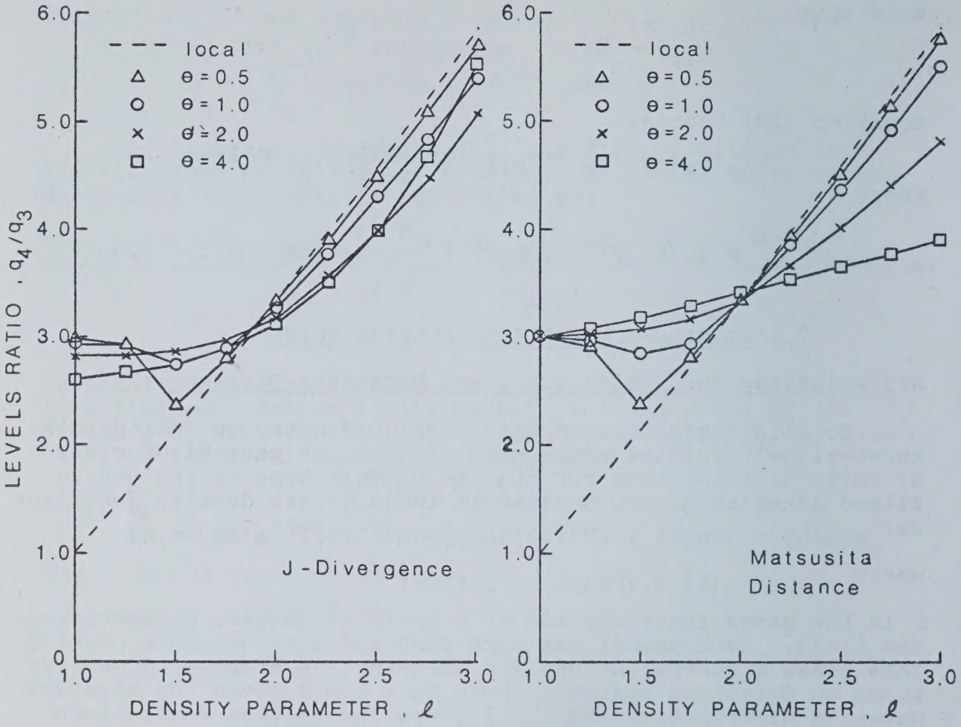


Figure 2: Levels Ratios for Maximum-Distance 4-Level Quantization

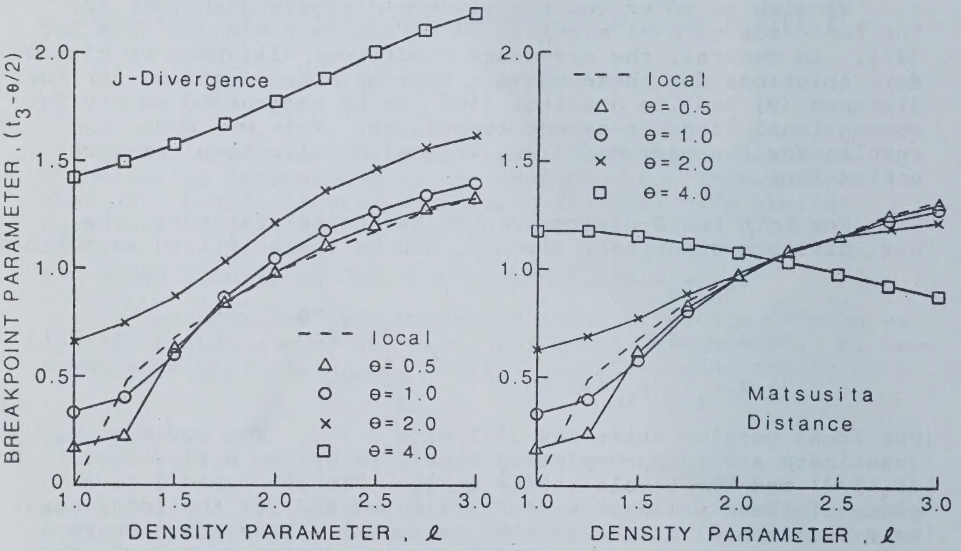


Figure 3: Breakpoint Parameters for Maximum-Distance 4-Level Quantization

B. Error Probabilities

Assuming H_0 and H_1 are equally likely, the quantizer-detector for (1) achieves minimum overall probability of error if we choose the threshold τ to be zero and the randomization a to be zero or unity. To compare the quantizers of Figs. 2 and 3, values of this probability of error are given in the Table for various values of ℓ, θ and sample size n.

Table: Error Probabilities for Quantizer-Detectors of Figs. 2 & 3
(M = Matsusita ; J = J-divergence)

a. density parameter $\ell = 1$

	$\theta = 1$			$\theta = 4$		
	n=5	n=10	n=20	n=5	n=10	n=20
M	$9.2\text{x}10^{-2}$	$3.3\text{x}10^{-2}$	$4.7\text{x}10^{-3}$	$7.0\text{x}10^{-5}$	$6.6\text{x}10^{-8}$	$4.1\text{x}10^{-14}$
J	$9.2\text{x}10^{-2}$	$3.3\text{x}10^{-2}$	$4.7\text{x}10^{-3}$	$7.7\text{x}10^{-5}$	$7.2\text{x}10^{-8}$	$5.0\text{x}10^{-14}$

b. density parameter $\ell = 2$

	$\theta = 1$			$\theta = 4$		
	n=5	n=10	n=20	n=5	n=10	n=20
M	$1.4\text{x}10^{-1}$	$6.9\text{x}10^{-2}$	$1.8\text{x}10^{-2}$	$1.1\text{x}10^{-5}$	$1.9\text{x}10^{-9}$	$3.4\text{x}10^{-17}$
J	$1.4\text{x}10^{-1}$	$7.0\text{x}10^{-2}$	$1.8\text{x}10^{-2}$	$2.1\text{x}10^{-5}$	$1.1\text{x}10^{-8}$	$8.2\text{x}10^{-16}$

c. density parameter $\ell = 3$

	$\theta = 1$			$\theta = 4$		
	n=5	n=10	n=20	n=5	n=10	n=20
M	$1.4\text{x}10^{-1}$	$6.6\text{x}10^{-2}$	$1.6\text{x}10^{-2}$	$1.6\text{x}10^{-6}$	$3.5\text{x}10^{-11}$	$1.2\text{x}10^{-20}$
J	$1.5\text{x}10^{-1}$	$6.8\text{x}10^{-2}$	$1.7\text{x}10^{-2}$	$1.2\text{x}10^{-5}$	$6.4\text{x}10^{-9}$	$2.3\text{x}10^{-16}$

VI. CONCLUSIONS

In this work we have studied the design of quantizer-detectors using the criterion of maximum distance. We have not considered the problem of choosing the appropriate distance measure; discussions of this problem can be found elsewhere [7]. Perhaps the most important result of this paper is that the local-case quantizer is independent of this choice. For the non-local case, the usefulness of these techniques is that they lead to a tractable formulation which can be solved by standard optimization techniques. These solutions are 'good' in some imprecise sense, and we may choose one that performs best.

ACKNOWLEDGEMENTS

This research is supported by the National Science Foundation under Grant-24187 and by the U.S. Army Research Office under Contract DAAG29-75-G-0192.

REFERENCES

1. Y.C. Ching and L. Kurz, "Nonparametric detectors based on m-interval partitioning," IEEE Trans. Inform. Theory, Vol. IT-18, pp. 251-257, March 1972.

2. W.J. Bushnell and L. Kurz, "The optimization and performance of detectors based on partition tests," Proc. Twelfth Annual Allerton Conf. Circuit and System Theory, pp. 1016-1023, Oct. 1974.

3. S.A. Kassam and J.B. Thomas, "Generalizations of the sign detector based on conditional tests," IEEE Trans. Commun., Vol. COM-24, pp. 481-487, May 1976.

4. T.S. Ferguson, Mathematical Statistics. New York: Academic, 1967.

5. H. Kobayashi and J.B. Thomas, "Distance measures and related criteria," Proc. Fifth Annual Allerton Conf. Circuit and System Theory, pp. 491-500, Oct. 1967.

6. S.M. Ali and S.D. Silvey, "A general class of coefficients of divergence of one distribution from another," J. Royal Stat. Soc., Series B, Vol. 28, pp. 131-142, 1966.

7. T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," IEEE Trans. Commun. Tech., Vol. COM-15,pp. 52-60, Feb. 1967.

A STOCHASTIC FRAMEWORK FOR OPTIMAL CURVE FITTING

HOWARD L. WEINERT
E. E. Dept.
Johns Hopkins University
Baltimore, MD 21218

GURSHARAN S. SIDHU
E. E. Dept.
SUNY Buffalo
Amherst, NY 14260

ABSTRACT

We show how to exploit stochastic smoothing techniques to develop dynamical recursive algorithms, and error bounds, for the deterministic (nonrandom) problem of spline interpolation (optimal curve fitting). We also present preliminary results on optimal sampling.

I. Introduction

A basic problem in numerical analysis is that of approximating the value of a function f at a point $t \in [0,T]$ given measurements $\lambda_1 f, \lambda_2 f, \ldots, \lambda_N f$, where $\{\lambda_j\}_1^N$ are linear functionals. Before a mathematical development can be considered, some additional structure must be placed on this problem. There are of course many ways to proceed, but we shall assume first that the function $f(\cdot)$ belongs to the Hilbert space H, where

$$H = \{g(t)\ ,\ t \in [0,T] : g^{(n)} \text{ exists a.e.}, \int_0^T [g^{(n)}]^2 < \infty\} \qquad (1)^{\dagger}$$

and n is a positive integer satisfying $1 \leq n \leq N$. Second, we shall assume that the measurement functionals are linearly independent and bounded on H. This assumption is really not restrictive since dependent and/or unbounded functionals provide no useful information. Next, let L be a differential operator:

$$L = D^n + a_{n-1}(t)\, D^{n-1} + \ldots + a_1(t)\, D + a_0(t) \qquad (2)$$
$$D \triangleq \frac{d}{dt}$$

and $a_j(\cdot)$, $0 \leq j \leq n-1$, has j continuous derivatives on [0,T]. Our third assumption will be that $\{\lambda_j\}_1^n$ are linearly independent on the null space of L. This condition, which is easily satisfied, guarantees a unique solution to the problem we are developing.

Now the measurements $\{\lambda_j f\}_1^N$ localize f to a linear variety (translated subspace) U(r) in H, where

$$U(r) = \{g \in H: \lambda_j g = r_j,\ 1 \leq j \leq N\} \qquad (3)$$
$$r_j \triangleq \lambda_j f$$

---

† $g^{(j)}$ denotes the j^{th} derivative of g.

Since there generally will be more that one function in U(r), we shall proceed by constructing a particular function s(•) ε U(r), and then approximating f(t) by s(t). The function s(•) that we shall choose is the Lg-spline interpolating $\{r_j\}_1^N$ with respect to $\{\lambda_j\}_1^N$ - namely, the function s(•) in U(r) that satisfies

$$\int_0^T (Ls)^2 = \min_{g \, \varepsilon \, U(r)} \int_0^T (Lg)^2 \tag{4}$$

Our assumptions guarantee that the optimization problem (4) has a unique solution. The Lg-spline is thus the smoothest function (in the sense of (4)) that interpolates the data. The superiority of splines over ordinary polynomials in interpolation and approximation is discussed in [1].

Spline functions have been extensively applied to many important problems in statistics and engineering (see [2] for a survey), including density and spectrum estimation, control, signal reconstruction, data compression, system identification, and nonlinear estimation. Real-time applications, and those in which data storage is a factor, have emphasized the need for recursive spline algorithms. The first recursive spline algorithms were described by Weinert [3] and Weinert and Kailath [4]. Their approach, however, does not exploit the dynamical structure of the spline problem. The present paper will expose and exploit this structure. This paper is an outline - only the main points will be presented, and without proofs. A full development of these results can be found in Weinert and Sidhu [5] and the other references listed at the end.

II. Lg-Splines and Stochastic Estimation

As we have seen, spline interpolation is entirely nonrandom. However, associated with every spline problem is a stochastic estimation problem which illuminates the structure of the spline problem and points the way to recursive spline algorithms.

The existence of this associated stochastic problem (or "randomized" spline problem) follows from two facts. First, H is a reproducing kernel Hilbert space. That is, there exists a function K(•,•) on [0,T] x [0,T] such that

$$\begin{gathered} K(\cdot,t) \; \varepsilon \; H \text{ for each } t \; \varepsilon \; [0,T] \\ \langle f(\cdot), K(\cdot,t)\rangle = f(t) \text{ for any } f \; \varepsilon \; H, \; t \; \varepsilon \; [0,T] \end{gathered} \tag{5}$$

where $< \cdot , \cdot >$ is the inner product in H. The function $K(\cdot,\cdot)$ is the reproducing kernel. Every reproducing kernel is also a covariance function; that is, there exists a zero-mean random process $y(t)$, $t \in [0,T]$, such that

$$E\, y(t)y(\tau) = K(t,\tau) \quad , \; t,\tau \in [0,T]. \tag{6}$$

Let Y be the Hilbert space of random variables spanned by $\{y(t),\ t \in [0,T]\}$, with inner product

$$< v,w >_{Y} = Evw,\ v,\ w \in Y. \tag{7}$$

Then, as first noted by Loève [6, p. 408], H and Y are congruent[††]; furthermore, $z \in Y$ and $g \in H$ are images under this congruence ($z \leftrightarrow g$) if and only if for each $t \in [0,T]$, $g(t) = E\, zy(t)$.

Using these facts, it can be shown that, with the proper choice of reproducing kernel, if $\hat{y}(t)$ is the linear least-squares estimate of $y(t)$ given random variables $\{\lambda_j y\}_1^N$, and if $\hat{\hat{y}}(t)$ is the value of $\hat{y}(t)$ obtained by setting $\lambda_j y = r_j$, $1 \leq j \leq N$, then

$$s(t) = \hat{\hat{y}}(t) \quad , \; t \in [0,T]. \tag{8}$$

Thus the Lg-spline interpolating $\{r_j\}_1^N$ with respect to $\{\lambda_j\}_1^N$ is a particular sample function of the random process $\hat{y}(t)$. Therefore, an algorithm for $\hat{y}(t)$ becomes an algorithm for $s(t)$ if $\lambda_j y$ is replaced by r_j.

There are many possible choices of inner product and reproducing kernel for the space H. However, there is only one choice which will permit us to derive Eq. (8)[†††]. We shall not give here the particular $K(\cdot,\cdot)$ that we need, since, unlike the results in [3]-[4], our algorithm does not require the explicit calculation of $K(\cdot,\cdot)$. This is an advantage since the calculation of $K(\cdot,\cdot)$ entails the solution of a multipoint boundary value problem.

III. Dynamical Model for y(•)

The random process $y(\cdot)$, with covariance equal to the particular reproducing kernel discussed above, is a lumped process; i.e., it is the output of finite-dimensional linear system driven by white noise. In particular

$$Ly(t) = u(t) \quad , \; t \in [0,T] \tag{9}$$

[††] Inner product spaces H_1 and H_2 are said to be congruent if there exists a one-to-one, onto linear mapping $\sigma: H_1 \rightarrow H_2$ that preserves inner products; i.e., $<g,h>_{H_1} = <\sigma(g), \sigma(h)>_{H_2}$, for all $g,h \in H_1$.

[†††] Other choices may also lead to randomized spline problems, but these estimation problems will be more complicated and may not be useful in developing recursive algorithms - see, for example, [7].

where L is given by (2) and $u(\cdot)$ is zero-mean white noise - $E\, u(t)u(\tau) = \delta(t-\tau)$. Writing (9) in state-variable form, we have

$$\begin{aligned} y(t) &= c\, x(t) \qquad t \in [0,T] \\ \frac{d}{dt}\, x(t) &= A(t)x(t)+bu(t) \end{aligned} \tag{10}$$

where

$$\begin{aligned} x'(\cdot) &= [y(\cdot) \quad y^{(1)}(\cdot) \ldots \ldots y^{(n-1)}(\cdot)\ ] \\ c &= [1 \quad 0 \quad . \ . \ . \quad 0] \\ b' &= [0 \quad 0 \quad . \ . \ . \quad 1] \end{aligned}$$

$$A(\cdot) = \left[\begin{array}{c|c} 0 & I \\ \hline -a_0(\cdot) & -a_1(\cdot) \ \ . \ . \ . \ \ -a_{n-1}(\cdot) \end{array}\right]$$

Note that the last row of $A(\cdot)$ contains the coefficients of L. We see that the choice of L, which is the first step in the spline interpolation problem (see (4)), determines the dynamical structure of the randomized spline problem. More will be said about the choice of L later. To complete the specification of the model (10), we must provide initial conditions - the statistics of $x(0)$ and the correlation between $x(0)$ and $u(\cdot)$. Only one set of initial conditions will force the covariance of the output of (10) to equal $K(\cdot,\cdot)$. They are

$$\begin{aligned} &Ex(0) = 0,\ Ex(0)x'(0) = M^{-1}[I + Q]\ M^{-T} \\ &Ex(0)u(t) = -M^{-1}\Delta(t)\ ,\ t \in [0,T] \end{aligned} \tag{11}$$ ††††

where

$$M= \begin{bmatrix} \lambda_1 c\phi(t,0) \\ . \\ . \\ . \\ \lambda_n c\phi(t,0) \end{bmatrix} \quad , \ \Delta(t) = \begin{bmatrix} \lambda_{1(\tau)}\, c\phi(\tau,t)_+ b \\ . \\ . \\ . \\ \lambda_{n(\tau)}\, c\phi(\tau,t)_+ b \end{bmatrix}$$

$$Q = \int_0^T \Delta(\tau)\Delta'(\tau)d\tau \quad , \quad \phi(\tau,t)_+ = \begin{cases} \phi(\tau,t) & , \ \tau \geq t \\ 0 & , \ \tau < t \end{cases}$$

---

†††† $M^{-T} = [M']^{-1}$, and parenthetical subscripts on functionals indicate the variable with respect to which they operate.

and $\phi(\cdot,\cdot)$ is the state transition matrix for A:

$$\frac{\partial}{\partial t}\phi(t,\tau) = A(t)\phi(t,\tau) \quad , \quad \phi(\tau,\tau) = I.$$

IV. Dynamical Recursive Spline Algorithm for EHB Data

Although we can go somewhat further without additional restrictions on $\{\lambda_j\}_1^N$, we shall restrict ourselves to the important class of functionals of the form

$$\lambda_j f = \sum_{k=1}^{n} \alpha_{j,k}\, f^{(k-1)}(t_j) \quad , \ j=1,2,\ldots,N \tag{12}$$

where the points $\{t_j\}_1^N$, known as knots, are ordered as $0 \leq t_1 < t_2 < \ldots < t_N \leq T$, and $\{\alpha_{j,k}\}$ are known real numbers. Such functionals are said to be of extended Hermite-Birkhoff (EHB) type.

Now let

$$c_j = [\alpha_{j,1} \quad \alpha_{j,2} \quad \cdots \quad \alpha_{j,n}]$$

and define

$$\tilde{M} = \begin{bmatrix} c_1\phi(t_1,t_n) \\ \cdot \\ \cdot \\ \cdot \\ c_{n-1}\phi(t_{n-1},t_n) \\ c_n \end{bmatrix} , \quad \tilde{\Delta}(t) = \begin{bmatrix} c_1\phi(t_1,t)_{-}b \\ \cdot \\ \cdot \\ \cdot \\ c_n\phi(t_n,t)_{-}b \end{bmatrix}$$

$$\tilde{Q} = \int_{t_1}^{t_N} \tilde{\Delta}(\tau)\tilde{\Delta}'(\tau)\,d\tau \quad , \quad \phi(t,\tau)_{-} = \begin{cases} \phi(t,\tau), & t \leq \tau \\ 0 \quad , & t > \tau \end{cases}$$

We are now ready to state the algorithm. It is derived by solving the linear least-squares smoothing problem for $y(\cdot)$, taking advantage of the known dynamical model generating $y(\cdot)$, and then substituting r_j for $\lambda_j y$. To proceed, we have

$$s(t) = c\,\hat{x}(t|N) \quad , \ t \in [0,T] \tag{13}$$

where the n x 1 vector $\hat{x}(t|N)$ is computed via the following steps:

Step A: Initialization: Compute $\tilde{M}$, $\tilde{\Delta}(t)$ for $t_1 \leq t \leq t_n$, and $\tilde{Q}$. Let $r_0' = [r_1 \ r_2 \ \ldots \ r_n]$ and compute

$$x_0 = \tilde{M}^{-1} r_0 \ , \ \Pi_0 = \tilde{M}^{-1}\tilde{Q}\tilde{M}^{-T}.$$

Step B: Forward pass for $t_n \le t \le t_N$: Compute the following quantities for $n \le j \le N-1$

$$e_{j+1} = r_{j+1} - c_{j+1}\,\hat{x}_{j+1|j}$$

$$R_{j+1} = c_{j+1}\,P_{j+1|j}\,c'_{j+1} \qquad \text{(a scalar)}$$

$$K_{j+1} = P_{j+1|j}\,c'_{j+1}$$

where $\hat{x}_{j+1|j}$ and $P_{j+1|j}$ are obtained from the following recursions, starting with $\hat{x}_{n|n} = x_0$ and $P_{n|n} = \Pi_0$: for each $j=n, n+1, \ldots, N-1$, starting with $\hat{x}(t_j|j) = \hat{x}_{j|j}$ and $P(t_j|j) = P_{j|j}$ integrate forwards up to t_{j+1} the differential equations

$$\frac{d}{dt}\,\hat{x}(t|j) = A(t)\,\hat{x}(t|j) \quad , \; t \ge t_j$$

$$\frac{d}{dt}\,P(t|j) = A(t)\,P(t|j) + P(t|j)\,A'(t) + bb' \quad , \; t \ge t_j$$

then take $\hat{x}_{j+1|j} = \hat{x}(t_{j+1}|j)$ and $P_{j+1|j} = P(t_{j+1}|j)$. Next compute

$$\hat{x}_{j+1|j+1} = \hat{x}_{j+1|j} + K_{j+1}\,R_{j+1}^{-1}\,e_{j+1}$$

$$P_{j+1|j+1} = [I - K_{j+1}\,R_{j+1}^{-1}\,c_{j+1}]\,P_{j+1|j}\,[I - K_{j+1}\,R_{j+1}^{-1}\,c_{j+1}]'$$

Step C: Backward pass for $t_n \le t \le t_N$: Let $\eta_{N,N} = c'_N\,R_N^{-1}\,e_N$. Starting with the value of $\hat{x}(t_N|N)$ found in Step B, compute $\hat{x}(t|N)$ by performing the following steps for $j=N-1, N-2, \ldots, n$: starting with $\hat{x}(t_{j+1}|N)$ integrate backwards the equation

$$\frac{d}{dt}\,\hat{x}(t|N) = A(t)\hat{x}(t|N) + bb'\mu_{j+1}(t|N), \quad t_j \le t \le t_{j+1}$$

where $\mu_{j+1}(t|N)$ is found by integrating backwards the equation

$$-\frac{d}{dt}\,\mu_{j+1}(t|N) = A'(t)\,\mu_{j+1}(t|N) \quad , \; \mu_{j+1}(t_{j+1}|N) = \eta_{j+1,N}$$

where for $n \le j \le N-2$,

$$\eta_{j+1,N} = \mu_{j+2}(t_{j+1}|N) + c'_{j+1}\,R_{j+1}^{-1}\,[e_{j+1} - K'_{j+1}\mu_{j+2}(t_{j+1}|N)]$$

Step D: Backward pass for $t_1 \le t \le t_n$: Starting with $\hat{x}(t_n|N)$ obtained from Step C , integrate backwards

$$\frac{d}{dt}\,\hat{x}(t|N) = A(t)\hat{x}(t|N) + b\tilde{\Lambda}'(t)\,\tilde{M}^{-T}\,\mu_{n+1}(t_n|N)$$

Step E: Backward pass for $t \leq t_1$: Starting with $\hat{x}(t_1|N)$ from Step D, integrate backwards

$$\frac{d}{dt} \hat{x}(t|N) = A(t)\hat{x}(t|N) \quad , \ t \leq t_1$$

Step F: Forward pass for $t \geq t_N$: Starting with $\hat{x}(t_N|N)$ from Step B, integrate forwards

$$\frac{d}{dt} \hat{x}(t|N) = A(t) \ \hat{x}(t|N) \quad , \ t \geq t_N.$$

V. Error Bounds for Spline Interpolation

We will now return to the case of arbitrary bounded linear functionals considered in Section I-III. In this section we will determine what, if anything, can be said about the interpolation error $|f(t)-s(t)|$ when all we know about f is that $f \in H$. Much work has been done on upper-bounding this error - see, e.g., Schultz [11] and Golomb and Weinberger [12]. All of these bounds have the general form

$$|f(t)-s(t)| \leq dq(f) \quad , \quad \text{fixed } t \in [0,T] \tag{14}$$

where d is a known constant and q(f) is some nonlinear functional of f; for example, q(f) may be the norm of f in H. Obviously, in order to evaluate this bound, we must have additional a priori information about f in the form of q(f). In most cases, this information is either completely lacking or is too imprecise to give a tight bound. It would be desirable to have a tight upper bound that can be computed using only the information that $f \in H$. Unfortunately, this cannot be done. The quantity $|f(t)-s(t)|$ can be arbitrarily large. All is not lost, however, because we can bound a normalized version of this error. Using the Schwartz inequality, the reproducing property, and the basic congruence between H and Y, we can show that

$$\frac{|f(t)-s(t)|}{\|f\|} \leq e(t) \tag{15}$$

where $\|\cdot\|$ is the norm in H determined by the specified inner product, and

$$e^2(t) = E(y(t)-\hat{y}(t))^2 \tag{16}$$

that is, $e^2(t)$ is the mean-square error in the randomized problem. Note that e(t) can be computed using just the information at hand since the choice of L in the spline problem determines the statistics of the randomized problem. Note also that e(t) is independent of the data $\{\lambda_j f\}_1^N$ in the spline problem and can thus be precomputed before measurements are made.

Recursions for computing e(t) using the dynamical model (10)-(11) will appear in [10].

VI. Optimal Sampling for Interpolation

In this section we shall restrict ourselves to the case in which $\lambda_j f = f(t_j)$, $1 \le j \le N$. In this case the Lg-spline reduces to what is called an L-spline. It is known that an L-spline s(•) interpolating $\{r_j\}_1^N$ at $\{t_j\}_1^N$ satisfies the following four conditions (L^a is the formal adjoint of L):

$$\begin{aligned} &L^a L\, s(t) = 0 \ , \ t_j < t < t_{j+1} \ , \ 1 \le j \le N-1 \\ &Ls(t) = 0 \ , \ 0 < t < t_1 \ , \ t_N < t < T \\ &s(t_j) = r_j \ , \ 1 \le j \le N \\ &s(\cdot) \in C^{2n-2}\,[0,T] \end{aligned} \tag{17}$$

We can now see that the choice of L determines the functional form of the spline. If $L=D^n$ we have a polynomial spline - a piecewise polynomial of degree 2n-1. In general, there is no rationale for the choice of L other than to produce a spline with a simple mathematical structure. Hence the popularity of polynomial splines. However, if something is known about the functional form of f, a particular L may give smaller interpolation error than others. Also, in some applications of splines, L may be fixed by the problem at hand - e.g., optimal tracking [13].

To return to the subject of this section, consider the following experimental design problem. Suppose we can choose N distinct points in [0,T] at which to sample an unknown function $f \in H$, after which the samples will be interpolated with a piecewise linear (L=D) spline. If we want to choose $\{t_j\}_1^N$ to give the best approximation to f over the entire interval [0,T], it makes sense (see(15)) to minimize $\int_0^T e^2(t)dt$. This of course is a nonlinear optimization problem, but, using standard results from optimization theory, it can be shown that this problem has a unique solution; and, furthermore, that the optimal sample points are

$$\begin{aligned} &t_1^* = \frac{T}{3N-1} \\ &t_{j+1}^* = t_j^* + \frac{3T}{3N-1} \ , \ 1 \le j \le N-1 \end{aligned} \tag{18}$$

Generalization to other types of L-splines is presently under consideration.

REFERENCES

[1] Greville, T.N.E., ed.(1969). Theory and Applications of Spline Functions, Academic Press, New York.

[2] de Figueiredo, R.J.P. (1974). A glimpse at splines and their applications, IEEE Circuits and Systems Newsletter, vol. 7, no. 4, pp.4-11.

[3] Weinert, H.L. (1972). A reproducing kernel Hilbert space approach to spline problems with applications in estimation and control, Ph.D. Thesis, Dept. of Electrical Engineering, Stanford University, Stanford, CA.

[4] Weinert, H.L. and Kailath, T. (1974). Stochastic interpretations and recursive algorithms for spline functions, Annals of Statistics, vol. 2, pp. 787-794.

[5] Weinert, H.L. and Sidhu, G.S. (1976). A stochastic framework for recursive computation of spline functions - Part I: Interpolating splines, submitted to IEEE Trans. Info. Theory.

[6] Loève, M., Fonctions aléatoires du second ordre, pp. 367-420 in Levy, P. (1965). Processus Stochastiques et Mouvement Brownien, second edition, Gauthier-Villars, Paris.

[7] Kimeldorf, G. and Wahba, G. (1970). Spline functions and stochastic processes, Sankhya, vol. 132, pp. 173-180.

[8] Sidhu, G.S. and Weinert, H.L. (1976). Dynamical recursive algorithms for Lg-spline interpolation of EHB data, submitted to Mathematics of Computation.

[9] Weinert, H.L. and Sidhu, G.S. (1976). On uniqueness conditions for optimal curve fitting, submitted to J. Optimization Theory and Appl.

[10] Weinert, H.L. and Sidhu, G.S. (1976). Stochastic error analysis of spline interpolation, in preparation.

[11] Schultz, M. (1973). Spline Analysis, Prentice-Hall, Inc., Englewood Cliffs, N.J.

[12] Golomb, M. and Weinberger, H. (1959). Optimal approximation and error bounds, On Numerical Approximation, ed. R. Langer, Univ. of Wisconsin Press, Madison, pp. 117-190.

[13] Weinert, H.L. and Kailath, T. (1976). A spline-theoretic approach to minimum-energy control, IEEE Trans. Automatic Control, vol. AC-21, pp. 391-393.

ON APPROXIMATIONS TO CONDITIONAL MEAN ESTIMATES

A. FOGEL AND S.C. SCHWARTZ
Dept. of Electrical Engineering and Computer Science
Princeton University
Princeton, New Jersey 08540

ABSTRACT

We consider approximations of conditional mean estimates under varying degrees of knowledge of the prior statistics. In this talk we focus on two estimates which are derived only from knowledge of the boundaries of the parameter space: a discrete maximum likelihood estimate (MLE) which has an exponential convergence property and a modified MLE which displays a min-max feature.

INTRODUCTION

Let θ represent an unknown parameter and X an observation drawn from the probability density function (pdf) $f(x|\theta)$. Conditional moments $E(x^j|\theta)$, $j=1,\ldots$, and posterior conditional moments of the form $E(\theta^j|x)$, when θ can be considered a random variable, are quantities of considerable interest in a variety of systems problems. These include multi-stage stochastic control ([1]) simultaneous detection and estimation ([2],[3]), adaptive communications ([4],[5]), and Bayes sequential probability ratio tests ([6]). In this paper we consider approximations to these conditional estimates under varying degrees of knowledge concerning the prior statistics. We shall focus on the conditional mean estimates (CME's), both the prior CME, $E(x|\theta)$, and the posterior version, $E(\theta|x)$. These estimates play a particularly important role in a broad class of detection problems ([7]) and in general system and estimation theory under a quadratic loss criterion.

One can initially ask the following type of question: what is the minimal amount of knowledge required to obtain good estimates of the parameter θ. Surprisingly, little really has to be known in order to obtain a sequence of asymptotically efficient estimates. A typical result is as follows. Let X_i be iid random variables drawn from an unknown distribution $F(x-\theta)$. Assume the pdf exists and is symmetric about the origin. Stone has shown ([8]) that there exists a sequence of estimates $\hat{\theta}_k$ such that the distribution of $(k^{1/2}(\hat{\theta}_k-\theta))$ converges to a zero-mean Gaussian pdf with variance equal to the inverse of the Fisher information contained in $\theta+x$. He calls his sequence adaptive maximum likelihood estimates and, in effect, he estimates the unknown distribution function F. A similar result is discussed by Sacks ([9]). His sequence of estimates is a linear combination of order statistics. Although both estimators are not simple, they do converge to the true value with minimal knowledge and assumptions on the underlying distribution.

One can also obtain consistent sequences of posterior CME's under minimal assumptions if one is again willing to construct estimates of pdf's. Consider the Bayes estimate (under a quadratic loss):

$$\begin{aligned}\hat{\theta}(x) &= \int \theta dP(\theta|x) \\ &= \int \theta \frac{f(x|\theta)}{f(x)} dG(\theta)\end{aligned} \tag{1.1}$$

where f(x) is the marginal pdf

$$f(x) = \int_{\theta} f(x|\theta)dG(\theta) \tag{1.2}$$

Assume for the moment that $f(x|\theta)$ is Gaussian with variance σ^2 and mean θ. Differentiating f(x) and rearranging, we obtain

$$\hat{\theta}(x) = \sigma^2 \frac{f'(x)}{f(x)} + x \tag{1.3}$$

It is clear that with no knowledge of the prior distribution $G(\theta)$, but knowing that $f(x|\theta)$ is Gaussian, one can obtain consistent estimates of θ by forming consistent estimates of the marginal pdf and its derivative. This example can immediately be generalized to an important class of pdf's, the exponential family,

$$f(x|\theta) = \exp(\theta x - b(\theta) + B(x)) \tag{1.4}$$

One can show, in a manner completely analogous to that above, that the posterior CME can be expressed as:

$$\hat{\theta}(x) = \frac{f'(x)}{f(x)} + B'(x) \tag{1.5}$$

Again, consistent estimates of the ratio f'/f lead to consistent estimates of the CME for θ with no knowledge of the prior pdf.

The above discussion should illustrate that with very little knowledge, one can obtain consistent estimates of the parameter θ if one is willing to accept a somewhat complicated estimator, i.e., forming estimates of pdf's or related quantities. Perhaps a more appropriate question, than the one posed earlier, is: how much knowledge is required in order to obtain consistent (or good) estimates which are, at the same time, relatively simple to implement.

We shall consider estimating the parameter θ under the assumption that the conditional pdf is a member of the exponential family. The conditional pdf is usually known in many problems, and the exponential family includes most of the pdf's commonly encountered in communication theory.[1] As to knowledge concerning the prior distribution, we shall assume only that the boundaries of the parameter space are known, i.e., we know that $\theta_l \leq \theta \leq \theta_u$.

In the next section we describe a modified MLE that displays a min-max property. In Section 3 we display a discrete maximum likelihood estimate (MLE) which has an interesting property of exponential convergence. Both these estimates can be viewed as being generated from the MLE. The motivation for this is related to the conditional density, (1.4).

Consider iid observations X_i, i=1,...k, governed by (1.4). The sample mean $k^{-1}t(\underline{x}) = k^{-1}\sum_{i=1}^{k} x_i$ is a sufficient statistic for θ and the conditional pdf for t can be written as

---

[1] This family, also known as the Koopman-Pitman-Darmois family, includes the Gaussian, gamma, beta, Poisson and binomial pdf's.

$$f(t|\theta) = \exp(\theta t-kb(\theta) + D(t)) \qquad (1.6)$$

The MLE of θ, denoted by $\tilde{\theta}_k$, satisfies

$$b'(\tilde{\theta}_k) = t/k = \bar{x} \qquad (1.7)$$

which we write as

$$\tilde{\theta}_k = \eta(t/k). = \eta(\bar{x}) \qquad (1.8)$$

Clearly, this estimate may be a nonlinear function of the data. In contrast, the CME $E(x|\theta) = b'(\theta)$ is also a quantity of interest The MLE for $b'(\theta)$ is the sample mean

$$\widetilde{b'}(\theta) = t/k \qquad (1.9)$$

which is a linear function of the data.

Katz has displayed admissible nonlinear estimates of $b'(\theta)$ by assuming an exponential prior on a truncated parameter space ([15]). He shows that the MLE (1.9), is inadmissible; however, it is not clear that the estimates proposed by Katz lead to a smaller Bayes risk. In addition, before resorting to nonlinear estimates, one should attempt to improve upon linear estimates, i.e., the MLE, and, since the MLE and CME are consistent, they coincide as $k\to\infty$. Consequently, any approximation of the CME must resemble (at least in the asymptotic sense) the MLE. That is why we have focused our initial efforts on linear modifications of the MLE. The goal is to derive estimates which result in a smaller Bayes risk than can be achieved with the usual MLE.

A MODIFIED MLE

In this section, we describe a modified MLE with a min-max property which utlizes knowledge of the boundary of the parameter space. Suppose that it is desired to estimate some function of θ, say $\mu(\theta)$, under a quadratic loss. The estimate is

$$\hat{\mu}_1 = \int \mu(\theta)dP(\theta|\underline{x}) \qquad (2.1)$$

and requires knowledge of the conditional pdf, $f(\underline{x}|\theta)$, and the prior distribution $G(\theta)$. Now consider a constrained estimate of the form

$$\hat{\mu}_2 = \sum_{i=1}^{N} \lambda_i h_i(\underline{x}) \qquad (2.2)$$

where the h_i are known functions of the data and λ_i are unknown constants. The Bayes risk is

$$\int_{\underline{x}} \int_{\theta} [\mu(\theta)-\hat{\mu}_2]^2 \; f(\underline{x}|\theta)dG(\theta)dx$$

Using least squares theory, Goldstein has displayed a class of estimates which require varying degrees of knowledge concerning the conditional pdf and the prior distribution. The amount of knowledge required clearly is related to the constraints and form imposed on the $h_i(x)$ ([10],[11]). These approximate Bayes solutions have been applied to communication problems in a variety of contexts; a recent exposition can be found in [12].

Take $\mu(\theta)=E(x|\theta)$ and assume that the h_i are constrained to be linear functions of the observations, i.e., one desires a linear Bayes estimate of the conditional mean $E(x|\theta)$. The estimate is then the usual linear combination of the sample mean and prior mean, weighted by the appropriate variances ([11]):

$$\hat{\mu}_3 = \frac{V(\mu)\overline{x} + V(\overline{x})m}{V(\mu)+V(\overline{x})} \tag{2.3}$$

where

$$m = \int \mu(\theta)dG(\theta) \tag{2.4}$$

$$V(\mu) = \int (\mu(\theta)-m)^2 dG(\theta) \tag{2.5}$$

$$V(\overline{x}) = \iint (\overline{x}-\mu(\theta))^2 f(\overline{x}|\theta)dG(\theta)dx \tag{2.6}$$

Thus knowledge of the prior mean (2.4) and the indicated variances is required. If the h_i are quadratic in the observations, higher order moments are needed, and so on.

We now consider a linear estimate of the form $\hat{\mu}(\theta)=\alpha\overline{x}$ under the assumption that the conditional pdf is exponential (see Eq. (1.6)) and that the parameter lies within a known parameter space $\theta_1 \le \theta \le \theta_u$. Recall that for the exponential family

$$\mu(\theta) = E(x|\theta) = b'(\theta), \tag{2.7}$$

so what we are after is a linear estimate of $b'(\theta)$. The MLE of $b'(\theta)$ is the sample mean

$$y_1 = \hat{\mu}(\theta) = \overline{x} = \frac{1}{k}\sum_{i=1}^{k} x_i \tag{2.8}$$

with the associated risk runction

$$R_1 = R(y_1, b'(\theta)) = b''(\theta)/k \tag{2.9}$$

Let the estimate $\alpha\overline{x}$ be denoted by y_2. The risk function is easily calculated to be

$$\begin{aligned} R_2 &= R(y_2, b'(\theta)) = E_\theta(\alpha\overline{x} - b'(\theta))^2 \\ &= \alpha^2 V_\theta(\overline{x}) + (\alpha-1)^2(b'(\theta))^2 \\ &= \frac{b''(\theta)}{k}\left[\alpha^2 + (\alpha-1)^2 k \frac{(b'(\theta))^2}{b''(\theta)}\right] \end{aligned} \tag{2.10}$$

It follows from the properties of the exponential family that $b(\theta)$ and its first two derivatives are continuous functions of θ. Hence, the quantity $(b'(\theta))^2/b''(\theta)$ achieves a (not necessarily unique) maximum in the compact domain, say at the point θ^*. Set

$$\beta(\theta^*) = \max_{\theta\epsilon[\theta_1,\theta_u]} \frac{(b'(\theta))^2}{b''(\theta)} \tag{2.11}$$

An upper bound on the risk function is obtained,

$$R_2 \le R_1[\alpha^2 + (\alpha-1)^2 k\,\beta(\theta^*)] \tag{2.12}$$

The coefficient which minimizes this bound is

$$\alpha_k = \frac{k\,\beta(\theta^*)}{1+k\,\beta(\theta^*)} = \frac{1}{1+\dfrac{1}{k\,\beta(\theta^*)}} \tag{2.13}$$

and upon substitution into (2.12), the upper bound becomes

$$R_2 \le \alpha_k\, R_1 \tag{2.14}$$

An inspection of the coefficient α_k shows that it is strictly less than one whenever $\beta(\theta^*) \neq \infty$. $\beta(\theta)$ is a non-negative quantity (see (2.11) and (2.9)) and is defined except for the degenerate cases $b''(\theta^*)=0$ and $b'(\theta^*)=\infty$, which we exclude by assumption. Thus, for all practical purposes, the estimate

$$y_2 = \alpha_k\,\overline{x} = \frac{1}{1+\dfrac{1}{k\,\beta(\theta^*)}}\;\frac{1}{k}\sum_{i=1} x_i \tag{2.15}$$

is a better estimate than the MLE in terms of the risk function. Hence, averaging over the unknown prior, it has a smaller Bayes risk. Note that $\alpha_k \to 1$ and the two estimates coincide asymptotically.

To recapitulate, we have assumed the boundaries of the unknown parameter as being given and have exploited an elementary property of the exponential family to improve upon the MLE estimate of $b'(\theta)$. This modified MLE can serve as an approximation to the CME $E(x|\theta)$. More formally, y_2 solves the problem

$$y_2 = \min_{\alpha}\,\max_{\beta}\, E_\theta\,\{(\alpha\,\overline{x} - b'(\theta))^2\}. \tag{2.16}$$

To illustrate the estimator, we consider two examples. Let $f(x|\theta)$ be Gaussian with unit variance. We have, $b(\theta) = \theta^2/2$, $\beta(\theta) = \theta^2$ and, with $|\theta_u|>|\theta_1|$, the coefficient is

$$\alpha_k = \frac{1}{1 + 1/k\,\theta_u^2} \tag{2.17}$$

where θ_u is the upper boundary. (For this case, since $b'(\theta)=\theta$, we are really discussing an estimate of the paramter θ which is better than the MLE.) With Poisson statistics, $P(x|\lambda)=\lambda^x e^{-\lambda}/x!$, we reparameterize by setting $\theta = \log\lambda$, i.e.,

$$P(x|\theta) = \exp[\theta x - e^\theta - \log x!] \tag{2.18}$$

Thus, $b(\theta) = \exp(\theta)$, $\beta(\theta) = e^\theta = \lambda$, and the estimate of $b'(\theta)$, which is the rate parameter, is

$$y_2 = \frac{1}{1+\dfrac{1}{k\lambda_u}}\;\overline{x} \tag{2.19}$$

A natural generalization of the estimate y_2 is to consider

$$y_3 = \alpha_k\overline{x} + (1-\alpha_k)\,b'(\theta_0), \tag{2.20}$$

where θ_0 is to be chosen. The risk function is

$$R_3 = E_\theta (y_3 - b'(\theta))^2$$

$$= \alpha_k{}^2 R_1 + (1-\alpha_k)^2 [b'(\theta) - b'(\theta_0)]^2 \qquad (2.21)$$

An alternate form, more useful for comparison purposes, is

$$R_3 = R_2 + (1-\alpha_k)^2 b'(\theta_0)[b'(\theta_0) - 2b'(\theta)] \qquad (2.22)$$

Since $V(x|\theta) = b''(\theta)>0$, $b'(\theta)$ is a strictly increasing function. Three cases need to be examined separately. For the first case $b'(\theta_u)>b'(\theta_1)>0$, and choosing $\theta_0=\theta_1$, it follows, with probability one, that the second term in (2.22) is negative and $R_3<R_2$. A similar conclusion follows for the case $b'(\theta_1)<b'(\theta_u)<0$ and upon choosing $\theta_0 = \theta_u$. The case $b'(\theta_1)<0 <b'(\theta_u)$ is more difficult. With some additional prior knowledge an improvement over y_2 can be made. For example, if we knew the expected value of $b'(\theta)$, an estimate of the form

$$y_4 = \alpha_k \overline{x} + (1-\alpha_k) E(b'(\theta)) \qquad (2.23)$$

can be shown to yield a smaller risk than that resulting from y_2.

It is interesting to compare this last estimate with that discussed by Goldstein. Equation (2.3) can be rewritten as

$$\hat{\mu}_3 = \beta_k \overline{x} + (1-\beta_k) E(\mu(\theta)) \qquad (2.24)$$

where

$$\beta_k = \frac{1}{1 + \frac{V(\overline{x})}{V(\mu)}} \qquad (2.25)$$

Recalling the definition of α_k (see Eqs. (2.11) and (2.13)), the similarity of the estimates $\hat{\mu}_3$ and y_4 is clear; the difference appears in the ratio terms, $b''(\theta^*)/(b'(\theta^*))^2$ compared to $E\{b''(\theta)\}/E\{(b'(\theta) - m)^2\}$. $\hat{\mu}_3$ is, by definition, the best linear estimate and y_4 requires less prior information.

DISCRETE MLE

When the parameter is defined on a compact domain, other variants of the MLE may be useful approximations to the CME. One such modification is the discrete MLE. Suppose the parameter can only take on one of n finite values, $\theta\epsilon\Theta=\{\theta_1,\ldots,\theta_n=\theta_u\}$. It is not difficult, nor is it surprising, to show that the discrete MLE, based on the k observations $\underline{x}$, converges exponentially fast to the true parameter. When the parameter can assume any value between θ_1 and θ_u, or, when it takes on a value outside the parameter space, the discrete MLE again exhibits an exponential convergence to the value $\theta_m\epsilon\Theta$. This limit point satisfies

$$\theta_m = \arg\{\max_{\theta_j \epsilon \Theta} E_{\theta_\nu} \log(f(x|\theta_j))\} \qquad (3.1)$$

where θ_ν is the true parameter. In other words, θ_m is that point in the discrete parameter space which maximizes the information function. These results, which will be described below, are related to a paper by Liporace ([13])[1] who developed Bayes CME's for a discrete parameter space.

Let the observations be iid and governed by the pdf $f(x_j|\theta_i)$, $j=1,\ldots k$, $i=1,\ldots,n$. Define the likelihood function

$$Z_k(\theta_i) = \frac{1}{k}\sum_{j=1}^{k} \log f(x_j|\theta_i) \tag{3.2}$$

With $\hat{\theta}(k)$ the discrete MLE,

$$\hat{\theta}(k) = \arg\{\max_{1\leq i\leq n} Z_k(\theta_i)\}, \tag{3.3}$$

and θ_ν the true value,

$$\begin{aligned}\sigma^2(k) &= E_{\theta_\nu}\{(\hat{\theta}(k) - \theta_\nu)^2\} \\ &= \sum_{i=1}^{n} (\theta_i-\theta_\nu)^2 P_{\theta_\nu}(\theta_i)\end{aligned} \tag{3.4}$$

where

$$P_{\theta_\nu}(\theta_i) = \mathrm{Prob}\{\max_j Z_k(\theta_j) = Z_k(\theta_i)|\theta_\nu\} \tag{3.5}$$

Note that $P_{\theta_\nu}(\theta_i)$, for $i\neq\nu$, is bounded from above by

$$P_e(k) = \Pr\{\max_j Z_k(\theta_j) > Z_k(\theta_\nu), j\neq\nu\} \tag{3.6}$$

which is the probability of making an error, i.e., an incorrect decision. Defining $R^2 = \max_j (\theta_j-\theta_\nu)^2$, we have the bound

$$\sigma^2(k) \leq (n-1)^2 R^2 P_e(k) \tag{3.7}$$

Obviously, the estimation problem reduces to the corresponding decision problem. Exponential convergence then follows for those detection problems where the underlying statistics lead to a probability of error which can be bounded by $P_e(k)\leq\gamma^k$, where $0<\gamma<1$, and k is the number of observations. For Gaussian conditional pdf, one has the bound

$$\sigma^2(k) < (2/k\pi r^2)^{1/2}(n-1)^2 e^{-kr^2/8} \tag{3.8}$$

where

$$r^2 = \min_{k\neq\nu} (\theta_k-\theta_\nu)^2. \tag{3.9}$$

In the case of a general pdf, we can apply the Chernoff upper bound in the form ([16])

$$P(B) \leq E\{\exp[\lambda(z-A)]\}\ ,\ \lambda \geq 0 \tag{3.10}$$

where the event B is defined as $B = \{x: z(x)>A\}$. Equation (3.7) becomes

---

[1] For a treatment of a similar problem in a dynamical system setting, see [14].

$$\sigma^2(k) \leq (n-1)^2 R^2 \max_{j \neq \nu} [\rho(j,\nu,\lambda)]^k \tag{3.11}$$

$$\rho(j,\nu,\lambda) = E_{\theta_\nu}\left[\frac{f(x|\theta_j)}{f(x|\theta_\nu)}\right]^\lambda < 1, j \neq \nu \tag{3.12}$$

and $\lambda \epsilon (0,1)$. Equality in the last expression is achieved iff $\theta_j = \theta_\nu$. We note that the exponential bound, in conjunction with the Borel-Cantelli lemma, demonstrates that the discrete MLE also converges with probability one.

For the situation where the true value is not contained in the discrete space, one can parallel the proof in Liporace to show that $\hat{\theta}(k)$ converges to θ_m which is defined in (3.1). An alternate proof involves expanding the function

$$E_{\theta_\nu}\left\{\left[\frac{f(x|\theta_j)}{f(x|\theta_m)}\right]^\lambda\right\} \tag{3.13}$$

in a Taylor series for λ. Note that this quantity involves θ_j, θ_ν, and θ_m, while the similar expression in (3.12) depends only on the first two. Under reasonable regularity conditions, it can be shown that the exponent λ can be chosen so that the above quantity is less than one. As a consequence, the mean square error

$$\sigma^2(k) = E_{\theta_\nu}(\hat{\theta}(k) - \theta_m)^2 \tag{3.14}$$

can be bounded by a quantity similar to (3.11).

Returning to the problem where the observations are Gaussian distributed, we note that the n likelihood functions, Eq. (3.2), need not be calculated. With $\overline{x}$ the sample mean, an inspection of the pdf shows that the discrete MLE is that point in the discrete space Θ closest to the sample mean, i.e., $Z_k(\theta_i)$ is proportional to $(\overline{x}-\theta_i)$. Hence, only the sample mean need be computed. Furthermore, with $\hat{\theta}(k)$ the discrete MLE, the estimate

$$\tilde{\theta}(k) = \frac{1}{k}\hat{\theta}(k) + (1 - \frac{1}{k})\overline{x} \tag{3.15}$$

can be shown to have a smaller risk function for $k > k_0$ and, of course, this estimate is equivalent to $\overline{x}$ as $k \to \infty$.

CONCLUSION

Two approximations for conditional mean estimates have been presented. Both estimates require the conditional pdf of the observations, $f(x|\theta)$. The only assumption concerning the prior statistics is knowledge of the boundaries of the parameter space.

For the first approximation, we assumed the conditional pdf was a member of the exponential family and exploited an elementary property. The estimate was a modified MLE which had a min-max property and the risk function was shown to be less than the corresponding risk for the MLE. Since the MLE for $E(x|\theta)$ is the sample mean, the estimate can also be viewed as a linear estimate with the coefficient chosen according to a min-max criterion. Another viewpoint may be adopted by writing the estimate in recursive form:

$$y_k = \alpha_k \overline{x} = y_{k-1} + \frac{\beta}{1 + \beta k} (x_k - y_{k-1})$$

This can be interpreted as a stochastic approximation procedure with the coefficient $\beta/(1+\beta k)$ satisfying the usual conditions. The quantity β, however, depends on the underlying pdf (see (2.11)). It would appear worthwhile to investigate similar stochastic approximation procedures where the coefficients are chosen according to a min-max risk function or similar criterion, and where some knowledge of the underlying statistical structure is taken into account.

The discrete MLE exhibits an exponential convergence to a point in the discrete space which maximizes the information between the point and the true value. It would appear that the discrete MLE, when combined with the usual MLE, will provide a better estimate than the MLE alone in terms of the risk function. This is indeed the case for Gaussian conditional pdf.

Although $f(x|\theta)$ has to be known in order to compute the discrete MLE, we have not made use of any particular property of the pdf. A continued investigation of the discrete MLE should attempt to incorporate properties of the pdf. It does not appear possible to improve upon the exponential convergence; however, an improvement in computing the constants appearing in the bound appears likely.

REFERENCES

1. Bar-Shalom, Y., and Tse, E., "Dual Effect, Certainty Equivalence, and Separation in Stochastic Control," IEEE Trans. Auto. Control, Vol. AC-19, No. 5, pp.494-500, Oct. 1974.

2. Middleton, D. and Esposito, R., "Simultaneous Optimum Detection and Estimation of Signals in Noise," IEEE Trans. Inform. Theory, Vol. IT-14, May 1968, pp.434-444.

3. Fredriksen, A., Middleton, D., and Vande Linde, D., "Simultaneous Signal Detection and Estimation Under Multiple Hypotheses," IEEE Transactions on Inform. Theory, Vol. IT-18, Sept. 1972, pp.607-614.

4. Birdsall, T.G., and Gobien, J.O., "Sufficient Statistics and Reproducing Densities in Simultaneous Sequential Detection and Estimation," IEEE Trans. Inform. Theory, Vol. IT-19, pp.760-768, Nov. 1973.

5. Schwartz, S., "On Detectors Determined by Estimators," 1975 Conference on Inform. Sciences and Systems, The Johns Hopkins University, Baltimore, Maryland, April, 1975, pp.307-312.

6. Blackwell, D., and Girshick, M.A., Theory of Games and Statistical Decisions, Wiley, 1954.

7. Schwartz, S.C., "Conditional Mean Estimates and Bayesian Hypothesis Testing," IEEE Trans. on Inform. Theory, Vol. IT-21, Nov. 1975, pp.663-665.

8. Stone, C., "Adaptive Maximum Likelihood Estimators of a Location Parameter," Ann. Math. Statist., Vol. 3, March 1975, pp.267-284.

9. Sacks, J., "An Asymptotically Efficient Sequence of Estimates of a Location Parameter," Ann. Math. Statist., Vol. 3, March 1975, pp.285-298.

10. Goldstein, M., "Approximate Bayes Solutions to Some Nonparametric Problems," Annals Math. Statist., Vol. 3, March 1975, pp.512-517.

11. Goldstein, M., "A Note on Some Bayesian Nonparametric Estimates," Ann. Statist., Vol. 3, 1975, pp.736-740.

12. Gardner, W.A., "Structurally Constrained Receivers for Signal Detection and Estimation," IEEE Trans. on Comm., Vol. COM-24, No. 6, pp.578-592, June 1976.

13. Liporace, L.A., "Variance of Bayes Estimates," IEEE Trans. on Inform. Theory, Vol. IT-17, November 1971, pp.665-669.

14. Hawkes, R.M., Moore, J.B., "Performance of Bayesian Parameter Estimators for Linear Signal Models," IEEE Trans. on Auto. Control, Vol. AC-21, Aug. 1976, pp.523-527.

15. Katz, M.W., "Admissible and Minimax Estimates of Parameters in Truncated Spaces," Ann. Math. Stat., Vol. 32, 1961, pp.136-142.

16. Chernoff, H., "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations," Annals Math. Statist., Vol. 23, 1952, pp.493-507.

ACKNOWLEDGEMENTS

The research reported here was supported by the National Science Foundation under Grant ENG-75-09610.

SIGNAL DETECTION BASED ON QUANTIZED DATA

SALEEM A. KASSAM
Department of Systems Engineering
Moore School of Electrical Engineering
University of Pennsylvania
Philadelphia, PA. 19174

ABSTRACT

Detection performance optimization with respect to input quantization is studied for weak-signal detection in additive noise. The optimum quantizer is shown to minimize the mean-squared-error between quantized data and its locally-optimum nonlinear function, when signal is absent. For non-Gaussian noise this quantization differs from the usual minimum-distortion quantization for the noise.

I. INTRODUCTION

The optimum quantization of data under minimum mean-squared-error and related criteria or under criteria based on entropy has been considered by various investigators [1-6]. Such investigations are concerned with the quantization of data in such a way as to loose the minimum amount of information in the quantization process, which may be necessary to meet constraints such as of storage capacity. When the data is primarily to be used to perform a test of statistical hypotheses it is clearly more appropriate to quantize in such a way as to obtain maximum discrimination of the outputs under the two hypotheses. Thus the optimum quantizer for detection of signals in noise can be expected to be generally quite different from the minimum-distortion quantizer for either signal-plus-noise or noise only.

In this paper we consider the detection of a weak signal in additive noise, for which a large number of data samples are generally required for adequate detection power. This is a situation in which some kind of quantization of the data would be very useful to cut down on system complexity. The optimum quantizers are derived for two common detection performance criteria appropriate for the weak-signal case, and are shown to be identical. More interestingly, it is shown that the optimum quantizer gives the best approximation, in the mean-squared-error sense, of the locally-optimum nonlinearity for the detection problem as specified by the generalized Neyman-Pearson Lemma [7].

We will study the following specific detection situation: given an observation vector $\underline{R} = (R_1, R_2, \ldots, R_n)$ of independent components, test the null hypothesis H_0 versus the alternative H_1, where the hypotheses are defined by the following equation

$$\begin{aligned} H_0&: \quad f_i(r) = f(r), \quad 1 \le i \le n, \quad \forall r \\ H_1&: \quad f_i(r) = f(r - \theta s_i), \quad 1 \le i \le n, \quad \forall r. \end{aligned} \tag{1}$$

Here the vector $\underline{s} = (s_1, s_2, \ldots, s_n)$ is a known signal vector, θ is a positive signal amplitude, $f(\cdot)$ is the density function of the additive noise, and $f_i(\cdot)$ is the density function of R_i. We assume that $f(\cdot)$ is symmetric about the origin and is absolutely continuous, and denote by $F(\cdot)$ the distribution function corresponding to the noise density function $f(\cdot)$. The objective is to specify the parameters of a (2m)-level symmetric quantizer so that detection performance based on the test statistic $T(\underline{R})$ given by

$$T(\underline{R}) = \sum_{i=1}^{n} s_i q(R_i; \underline{a}_m, \underline{\ell}_m) \tag{2}$$

is optimized. In (2) $q(\cdot; \underline{a}_m, \underline{\ell}_m)$ is the characteristic (staircase function) of a (2m)-level quantizer parameterized by m-vectors $\underline{a}_m = (a_1, a_2, \ldots, a_m)$ and $\underline{\ell}_m = (\ell_1, \ell_2, \ldots, \ell_m)$. The components of $\underline{a}_m$ satisfy the condition $a_j > a_{j+1}$ and $a_m = 0$, and form a partition of $[0, \infty)$. Defining a_0 to equal $+\infty$, and $\ell_{-j} = -\ell_j$, we have

$$q(R_i; \underline{a}_m, \underline{\ell}_m) = \begin{cases} \ell_j & R_i \varepsilon [a_j, a_{j-1}) \\ \ell_{-j}, & R_i \varepsilon (-a_{j-1}, -a_j] \end{cases} \tag{3}$$

(The ambiguity when $R_i = 0$ can be resolved by arbitrarily assigning $q(0; \underline{a}_m, \underline{\ell}_m) = 0$.)

II. DERIVATION OF THE OPTIMUM QUANTIZER

The optimum quantizer maximizing detection performance in testing H_0 vs. H_1 in the weak-signal large-sample case ($\theta \to 0$, $n \to \infty$) can be derived for suitable detection performance criteria. We now consider two common criteria.

(a) The Criterion of Local Optimality

Assume that the intervals partitioning the quantizer input are fixed by the parameter vector $\underline{a}_m$. In order to set the output levels ℓ_j we may employ the local-optimality detection criterion [7]; that is, we could require the ℓ_j to be such as to result in a locally-optimum test statistic for the detection problem specified by H_0 and H_1. A locally-optimum detector maximizes the slope of the detector power function at $\theta=0$, and is an appropriate criterion for the detection of weak signals.

We define the random variable Y_{ij} as

$$Y_{ij} = \begin{cases} 1, & R_i \varepsilon I_j \\ 0, & \text{otherwise} \end{cases} \qquad \text{for} \quad \begin{array}{l} 1 \le i \le n \;, \\ -m \le j \le m \;, \\ (j \neq 0) \end{array} \tag{4}$$

where $I_k = [a_k, a_{k-1})$ and $I_{-k} = (-a_{k-1}, -a_k]$, so that (2) can be written as

$$T(\underline{R}) = \sum_{i=1}^{n} s_i \sum_{\substack{j=-m \\ j\neq 0}}^{m} \ell_j \, Y_{ij} \; . \tag{5}$$

Consider now the joint distribution of the matrix $\underline{Y}_{ij}$ of random variables Y_{ij}, with $1 \le i \le n$ and $-m \le j \le m$, $j \neq 0$. It is easily seen that the joint density $p(\underline{y}_{ij})$ of $\underline{Y}_{ij}$ is given by

$$p(\underline{y}_{ij}) = P\{Y_{ij} = y_{ij}, \quad 1 \le i \le n \text{ and } -m \le j \le m, \quad j \neq 0\}$$

$$= \prod_{i=1}^{n} \prod_{\substack{j=-m \\ j\neq 0}}^{m} [p_{ij}(\theta)]^{y_{ij}} \; , \tag{6}$$

where

$$p_{ij}(\theta) = \begin{cases} F(a_{j-1} - \theta s_i) - F(a_j - \theta s_i) \, , & j > 0 \\ F(-a_{-j} - \theta s_i) - F(-a_{-j-1} - \theta s_i) \, , & j < 0 \end{cases} \tag{7}$$

provided that for each i there is only one integer j between $-m$ and m, $j \neq 0$, for which $y_{ij}=1$, with the rest being zero; otherwise $p(\underline{y}_{ij})=0$. For the locally-optimum test based on the $\{Y_{ij}\}$ the test statistic of the threshold test is given by [7]

$$T^*(\underline{R}) = \frac{d}{d\theta} \log p(\underline{Y}_{ij}) \Big|_{\theta=0}$$

$$= \sum_{i=1}^{n} \sum_{\substack{j=-m \\ j\neq 0}}^{m} Y_{ij} \frac{d}{d\theta} \log p_{ij}(\theta) \Big|_{\theta=0}$$

$$= \sum_{i=1}^{n} s_i \sum_{\substack{j=-m \\ m\neq 0}}^{m} \ell_j^* Y_{ij} \;, \tag{8}$$

where ℓ_j^* is given by

$$\ell_j^* = \frac{f(a_j)-f(a_{j-1})}{F(a_{j-1})-F(a_j)} \;, \quad j=1,2,\ldots,m \;. \tag{9}$$

and $\ell_{-j}^* = -\ell_j^*$.

Thus we conclude that, for a given $\underline{a}_m$ vector, the symmetric quantizer with output levels given by (9) gives the locally-optimum detector based on (2m)-level symmetrically quantized data. Note that this result holds for finite n.

It can be shown that the statistic $T(\underline{R})$ is asymptotically normally distributed, by considering its moment-generating function, assuming some mild conditions hold for $\underline{s}$. Thus the asymptotic power function of the test based on $T(\underline{R})$ is obtained from the normal distribution function, in terms of its mean and variance [8]. This consideration leads to the result that local power is maximized with respect to $\underline{a}_m$ by maximizing the ratio

$$E_{\ell^*,a} = \frac{d}{d\theta} E\{T^*(\underline{R})\}\Big|_{\theta=0} \Big/ \sqrt{\mathrm{Var}_0\{T^*(\underline{R})\}} \;. \tag{10}$$

(This follows because $[T^*(\underline{R}) - E\{T^*(\underline{R})\}] \big/ \sqrt{\mathrm{Var}\{T^*(\underline{R})\}}$ is asymptotically standard normal.) Since we have

$$E\{Y_{ij}\} = \begin{cases} F(a_{j-1} - \theta s_i) - F(a_j - \theta s_i) \;, & j>0 \\ F(-a_{-j} - \theta s_i) - F(-a_{-j-1} - \theta s_i) \;, & j<0 \end{cases} \tag{11}$$

and

$$\mathrm{Var}_0\{Y_{ij}\} = \begin{cases} F(a_{j-1}) - F(a_j) \;, & j > 0 \\ F(-a_{-j}) - F(-a_{-j-1}) \;, & j < 0 \;, \end{cases} \tag{12}$$

we obtain

$$E^2_{\ell^*,a} = 2 \sum_{i=1}^{n} s_i^2 \sum_{j=1}^{m} \frac{[f(a_j) - f(a_{j-1})]^2}{F(a_{j-1}) - F(a_j)} \;. \tag{13}$$

Now ignoring the term $\sum_{i=1}^{n} s_i^2$ in (13) (assume that $\sum_{i=1}^{n} s_i^2$ converges to a finite limit as $n\to\infty$), we find that the optimum parameter vector $\underline{a}_m = \underline{a}_m^*$ maximizes

$$G_{\ell^*,a} = 2 \sum_{j=1}^{m} \frac{[f(a_j) - f(a_{j-1})]^2}{F(a_{j-1}) - F(a_j)} \quad . \tag{14}$$

This expression can now be maximized with respect to $\underline{a}_m$ to complete the specification of the quantizer under the local-optimality criterion. At this point we consider a second criterion for optimization, and later will deduce the equations for the optimum $\underline{a}_m$ maximizing (14).

(b) The Criterion of Maximum Efficacy

The efficacy F of a threshold test for H_0 versus H_1 based on a test statistic T is defined as [7]

$$F = \lim_{n\to\infty} \frac{1}{n} \frac{\left[\frac{d}{d\theta} E_\theta\{T\}\big|_{\theta=0}\right]^2}{\mathrm{Var}_0\{T\}} \quad . \tag{15}$$

It is an asymptotic measure of the effectiveness of the test in detecting small signals, and may be described as an incremental-signal-to-noise ratio for the test statistic. Using efficacy as a measure of performance, we may look for the optimum (2m)-level quantizer maximizing efficacy.

For the statistic of (2), using (5), we get

$$E_\theta\{T(\underline{R})\} = \sum_{i=1}^{n} s_i \sum_{j=1}^{m} \{\ell_j[F(a_{j-1}-\theta s_i) - F(a_j-\theta s_i)]$$

$$+\ell_{-j}[F(-a_j-\theta s_i) - F(-a_{j-1}-\theta s_i)]\} \tag{16}$$

and

$$\frac{d}{d\theta} E_\theta\{S\}\big|_{\theta=0} = 2 \sum_{i=1}^{n} s_i^2 \sum_{j=1}^{m} \ell_j[f(a_j) - f(a_{j-1})]. \tag{17}$$

We also have

$$\mathrm{Var}_0\{T(\underline{R})\} = 2 \sum_{i=1}^{n} s_i^2 \sum_{j=1}^{m} \ell_j^2[F(a_{j-1}) - F(a_j)] \tag{18}$$

Thus the efficacy F is, from (15),

$$F = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} s_i^{\,2} \; \frac{2\left\{ \sum_{j=1}^{m} \ell_j [f(a_j) - f(a_{j-1})] \right\}^2}{\sum_{j=1}^{m} \ell_j^{\,2} [F(a_{j-1}) - F(a_j)]} \tag{19}$$

Assuming that $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} s_i^{\,2}$ exists and is finite (as in the constant-signal case when $s_i = s$, all i), and normalizing the efficacy by dividing by this limit, we obtain the equivalent objective of maximizing the normalized efficacy

$$G_{\ell,a} = \frac{2\left\{ \sum_{j=1}^{m} \ell_j [f(a_j) - f(a_{j-1})] \right\}^2}{\sum_{j=1}^{m} \ell_j^{\,2} [F(a_{j-1}) - F(a_j)]} \; . \tag{20}$$

In order to maximize $G_{\ell,a}$ we can proceed to first maximize it with respect to $\underline{\ell}_m$ for a given vector $\underline{a}_m$; setting the partial derivatives equal to zero, we find that the optimum values $\ell_j^{\,*}$ of the levels ℓ_j are given precisely by (9). With these optimum values for the ℓ_j substituted in (20), $G_{\ell,a}$ becomes $G_{\ell^*,a}$ given by (14). Thus we conclude that the optimum quantizers under the local-optimality and maximum-efficacy criteria are identical.

It should be noted that equations similar to (9) and (14) have been derived previously in related work [9].

III. QUANTIZATION OF THE OPTIMUM DETECTOR NONLINEARITY

The locally-optimum detector for H_0 vs. H_1 based on the unquantized observation vector $\underline{R}$ uses the test statistic

$$T_\ell(\underline{R}) = \sum_{i=1}^{n} s_i g(R_i), \tag{21}$$

where $g(\cdot)$ is given by [7]

$$g(x) = -\frac{d}{dx} \ln f(x) \;. \tag{22}$$

We now show a direct relationship between $g(x)$ and $q(x;\underline{a}_m^*,\underline{\ell}_m^*)$. Let us consider the mean-squared-error $C_{\ell,a}$ between $g(R_i)$ and $q(R_i;\underline{a}_m,\underline{\ell}_m)$ under the null hypothesis H_0; we have

$$C_{\ell,a} = E_0\{[q(R_i;\underline{a}_m,\underline{\ell}_m)]^2\} + 2E_0\{q(R_i;\underline{a}_m,\underline{\ell}_m)g(R_i)\}$$
$$+ I(f) \;, \tag{23}$$

where $I(f)$ is Fisher's Information of $f(\cdot)$, assumed to be finite. Note that $I(f)$ is independent of the quantizer parameters. The other two terms on the right side of (23) can be written explicitly as

$$E_0\{[q(R_i;\underline{a}_m,\underline{\ell}_m)]^2\} = 2 \sum_{j=1}^{m} \ell_j^2 \int_{a_j}^{a_{j-1}} f(x)\, dx \tag{24}$$

and

$$E_0\{q(R_i;\underline{a}_m,\underline{\ell}_m)g(R_i)\} = 4 \sum_{j=1}^{m} \ell_j \int_{a_j}^{a_{j-1}} df(x) \;. \tag{25}$$

For a given input partition vector $\underline{a}_m$, the mean-squared error $C_{\ell,a}$ can be minimized with respect to $\underline{\ell}_m$ by requiring that the first partial derivatives with respect to the ℓ_j on the right side of (23) be equal to zero. It is easily shown that the optimum values ℓ_j^* of the ℓ_j satisfy the previous equation (9) for the optimum ℓ_j. With these optimum values, the minimized mean-squared error $C_{\ell^*,a}$ is obtained directly to be

$$C_{\ell^*,a} = I(f) - 2 \sum_{j=1}^{m} \frac{[f(a_j) - f(a_{j-1})]^2}{F(a_{j-1}) - F(a_j)}$$

$$= I(f) - G_{\ell^*,a} \quad , \qquad (26)$$

where $G_{\ell^*,a}$ was defined in (14). Thus the further minimization of $C_{\ell^*,a}$ with respect to $\underline{a}_m$ is equivalent to maximization of $G_{\ell^*,a}$. We conclude, therefore, that the optimum quantizer derived in Section II is also the best approximation of the locally-optimum nonlinearity $g(\cdot)$, in the mean-squared-error sense.

We already have in (9) the specification of the optimum $\underline{\ell}_m$ vector for a given vector $\underline{a}_m$. We now obtain the equation for the optimum $\underline{a}_m$. Setting the partial derivatives of $C_{\ell,a}$ of (23) with respect to the a_j equal to zero, using (24) and (25), we find that

$$- \ell_j^2 f(a_j) + \ell_{j+1}^2 f(a_j) - 2\ell_j f'(a_j) + 2\ell_{j+1} f'(a_j) = 0,$$

$$j = 1,2,\ldots,m-1$$

so that

$$- \frac{f'(a_j)}{f(a_j)} = \frac{\ell_j + \ell_{j+1}}{2} \quad j = 1,2,\ldots, m-1 \ . \qquad (27)$$

Equation (27) together with the previously obtained

$$\ell_j = \frac{f(a_j) - f(a_{j-1})}{F(a_{j-1}) - F(a_j)} \ , \quad j = 1,2,\ldots,m \ , \qquad (28)$$

can therefore be solved to obtain the parameters of the optimum quantizer.

IV. CONCLUSION

A comparison with (27) and (28) of the equations for the parameters of the minimum mean-squared-error quantizer derived in [1] reveals that the optimum quantizer for the detection problem coincides with the minimum mean-squared-error quantizer (for the received data under H_0) only when $f(\cdot)$ is a Gaussian density. Thus when the additive noise is non-Gaussian there can be a considerable difference in the characteristics of the two optimum quantizers. As an interesting example, when $f(\cdot)$ is the double-exponential density the optimum quantizer for detection is simply the hard-limiter, giving rise to the sign detector. The hard-limiter is also the locally-optimum nonlinearity for this detection problem.

REFERENCES

1. J. Max, "Quantizing for minimum distortion," IRE Trans. Inform. Theory, vol. IT-6, pp. 7-12, March 1960.

2. A. I. Velichkin, "Optimum characteristics of quantizers," Telecommun. Radio Eng. (USSR), vol. 18, part II, pp. 1-7, Feb. 1963 (translated from Radiotechnika).

3. T. J. Goblick, Jr., and J. L. Holsinger," Analog source digitization: a comparison of theory and practice," IEEE Trans. Inform. Theory, vol. IT-13, pp. 323-326, April 1967.

4. D. G. Messerschmitt, "Quantizing for maximum output entropy," IEEE Trans. Inform. Theory, vol. IT-17, p. 612, Sept. 1971.

5. D. J. Goodman and A. Gersho, "Theory of an adaptive quantizer," IEEE Trans. Commun., vol. COM-22, pp. 1037-1045, Aug. 1974.

6. J. M. Morris and V. D. VandeLinde, "Robust quantization of discrete-time signals with independent samples," IEEE Trans. Commun., vol. COM-22, pp. 1897-1902, Dec. 1974.

7. J. Capon, "On the asymptotic efficiency of locally optimum detectors," IRE Trans. Inform. Theory, vol. IT-7, pp. 67-71, April 1961.

8. S. A. Kassam and J. B. Thomas, "Asymptotically robust detection of a known signal in contaminated non-Gaussian noise," IEEE Trans. Inform. Theory, vol. IT-22, pp. 22-26, January 1976.

9. S. A. Kassam and J. B. Thomas, "Generalizations of the sign detector based on conditional tests," IEEE Trans. Commun., vol. COM-24, pp. 481-487, May 1976.

MAXIMUM LIKELIHOOD SEQUENCE ESTIMATION FOR RANDOMLY DISPERSIVE CHANNELS†

ROBERT E. MORLEY, JR. and DONALD L. SNYDER
Washington University, St. Louis, Missouri 63130

ABSTRACT

In this paper, we combine the generalized likelihood-ratio formula, state-variable estimation theory, and the Viterbi algorithm to develop receiver designs for maximum likelihood sequence estimation on randomly dispersive channels. The results apply both to uncoded and convolutionally encoded systems that employ a variety of modulation formats, such as binary- or quadri-phase shift keying, frequency-shift keying, and amplitude-shift keying. For the model

$$dy(t) = s(t,i_1,i_2,\dots,i_N)dt + dw(t)$$

in which $i_j \in \{0,1,\dots,2^m-1\}$ for an m-ary information sequence, $w(t)$ is a Wiener process, and where there holds

$$E\{s(t,i_1,i_2,\dots i_N)|\mathcal{Y}_{0,t},i_1,i_2,\dots,i_N\}$$
$$= E\{s(t,i_1,i_2,\dots,i_N)|\mathcal{Y}_{\max(0,t-LT)},i_{\max(1,n-L)},\dots,i_n\}, \quad (1)$$

the sequence $\hat{i}_1,\dots,\hat{i}_N$ that maximizes the likelihood of observing a given realization of $y(t)$ on the interval $[0,NT]$ can be determined by employing a set of m^{L+1} estimators and a Viterbi algorithm having m^L states.

For a point process observation where points are observed with rate $s(t,i_1,\dots,i_N)$, where (1) holds, a similarly structured receiver can be used to determine the maximum likelihood sequence $\hat{i}_1,\dots,\hat{i}_N$.

Our result for maximum likelihood sequence estimation suggests receiver designs for a wide variety of randomly dispersive channels subject to the condition (1). These channels include Rayleigh and Ricean fading channels, random phase channels, randomly time-dispersive channels, and point process channels that occur in optical communication systems operating in atmospheric turbulence.

---

†This work was supported in part by the National Science Foundation under Research Grants ENG74-07800 and ENG76-11565 and by the National Institutes of Health under Research Grant RR00396 from the Division of Research Resources.

COMPARATIVE SENSITIVITY AND ABSOLUTE INVARIANT COMPENSATORS WITH APPLICATION TO A DC-AC CONVERTER*

R.M. DE SANTIS
Département de Génie Electrique
Ecole Polytechnique
Montréal, Québec

SERGE LEFEBVRE
Département de Génie Electrique
Ecole Polytechnique
Montréal, Québec

ABSTRACT

The present paper investigates the possibility of extending the classical Nyquist plot design procedures to the design of multivariable feedback compensators. The design objectives are stability, integrity and low sensitivity to external perturbations. The novelty of the approach is in the unveiling of a number of relevant interconnections between properties of the classical invariant compensators and the sensitivity matrix considered in more recent sensitivity studies.

1. INTRODUCTION

Let us consider a multivariate linear and time invariant system of the type illustrated in Figure 1 where, using Laplace transforms, the output vector y_o is a function of the input vector u and of the vector disturbance η according to the following equations

$$y_o = y^o + \Delta^o;\ y^o = [G]u;\ \Delta^o = [F]\eta \qquad (1)$$

In these equations [G] and [F] are matrix transfer functions, y^o denotes the nominal output, while Δ^o represents the influence of the disturbance over the output. To reduce the influence of the disturbance, the system in Figure 1 is often embedded in a feedback configuration of the type represented in Figure 2. In this case the system is described by the following equations

$$y = y^o + \Delta;\ \Delta = -[G]\,[M]\,\Delta + [F]\eta;\ y^o = [G]u \qquad (2)$$

From here one obtains $\Delta = \$\Delta^o$, where the matrix transfer function of the sensitivity operator \$ is given by [1]

$$[\$] = [I + [G][M]]^{-1} \qquad (3)$$

In what follows we will investigate how the classical Nyquist plot procedures which are usually confined to the scalar case can be extended to the design of [M] in the multivariate case. The design objectives are stability, stability under loop failure conditions, and sensitivity reduction.

---

* This research was supported in part by the Canadian National Research Council, grant CNRC-A8244, and by the Government of Quebec, grant FCAC-CRP-378-76.

We will proceed by developing and exploiting some relevant interconnections between the absolute invariant compensators considered in the classical literature, [2], and the above sensitivity operator matrix. These interconnections will lead to a multivariate feedback design approach which retains a geometrical visualization and a physical perception of the compensators' action similar to those which characterize the scalar case. The progress of the theory will be illustrated by the design of a DC-AC converter.

For other studies on the subject the reader is addressed to [3] and [4]. Though somewhat similar in their objectives these references use, however, an approach which is completely different from that presented here. This difference is due in part to the fact that while in [3] and [4] the emphasis is on reducing intercoupling, this design objective plays no role in the present development.

2. ABSOLUTE INVARIANT COMPENSATORS AND SENSITIVITY

For simplicity of exposition we will treat the case where the input and output vectors have dimension two, while the disturbance is given by a scalar quantity. In this case the configurations of Figures 1 and 2 become those of Figure 3 and 4. Equations (1), (2), and (3) become the following:

$$\begin{bmatrix} y_{o1} \\ y_{o2} \end{bmatrix} = \begin{bmatrix} \overset{o}{y}_1 \\ \overset{o}{y}_2 \end{bmatrix} + \begin{bmatrix} \overset{o}{\Delta}_1 \\ \overset{o}{\Delta}_2 \end{bmatrix} ; \begin{bmatrix} \overset{o}{y}_1 \\ \overset{o}{y}_2 \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} ; \begin{bmatrix} \overset{o}{\Delta}_1 \\ \overset{o}{\Delta}_2 \end{bmatrix} = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} \eta$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \overset{o}{y}_1 \\ \overset{o}{y}_2 \end{bmatrix} + \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix} ; \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix} = - \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix} + \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} \eta$$

$$\begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix} = [\$] \begin{bmatrix} \overset{o}{\Delta}_1 \\ \overset{o}{\Delta}_2 \end{bmatrix} ; \quad [\$] = \Big[I + [G][M] \Big]^{-1}.$$

The key definitions and properties of the absolute invariant compensators are as follows.

Definition 1. The absolute invariant, AI, compensator D^{11}, (D^{22}), is a scalar compensator with the property that if it is connected between the first output, (second output), and the first input, (second input), then it reduces to zero the influence of the disturbance over the second output (first input).

Definition 2. The AI compensator D^{12}, (D^{21}), is a scalar compensator with the property that if it is connected between the second output, (first output), and the first input, (second input), then it reduces to zero the influence of the disturbance over the first output (second output).

Proposition 1. (Analytic expression of the Absolute Invariant compensators). The AI compensators associated with the open loop system are given by the following expressions

$$D^{11} = \frac{F_2}{G_{21} F_1 - F_2 G_{11}} \qquad D^{21} = \frac{F_2}{G_{22} F_1 - F_2 G_{12}}$$

$$D^{22} = \frac{F_1}{G_{12} F_2 - F_1 G_{22}} \qquad D^{12} = \frac{F_1}{F_2 G_{11} - F_1 G_{21}}$$

Proposition 2. (Feedback structural properties of the absolute invariant compensators). The AI compensators D_c^{ij} associated with the closed loop system can be expressed in terms of the AI compensators D^{ij} associated with the open loop system according to the following formula

$$\begin{bmatrix} D_c^{11} & D_c^{12} \\ D_c^{21} & D_c^{22} \end{bmatrix} = \begin{bmatrix} D^{11} & D^{12} \\ D^{21} & D^{22} \end{bmatrix} \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}$$

where

$$\alpha = \left(1 - \frac{M_{11}}{D^{11}} - \frac{M_{21}}{D^{21}}\right), \quad \beta = \left(1 - \frac{M_{22}}{D^{22}} - \frac{M_{12}}{D^{12}}\right)$$

Proposition 3. (On the relation between the sensitivity operator and the Absolute Invariant compensators). The sensitivity operator matrix is given by the following formula

$$\$ = \frac{1}{K} \begin{bmatrix} \dfrac{D_c^{22}}{D^{22}} & 0 \\ 0 & \dfrac{D_c^{11}}{D^{11}} \end{bmatrix} = \frac{1}{K} \begin{bmatrix} \beta & 0 \\ 0 & \alpha \end{bmatrix}$$

where K is the product of the return difference gains of the basic feedback loops created by M_{ij}

$$K = (1 + G_{21}M_{12})\ (1 + G_{12}(12)M_{21})\ (1 + G_{22}(21,12)M_{22})\ (1 + G_{11}(22,21,12)M_{11})$$

Note also that $K = \det[I + [G]\ [M]]$.

Proposition 4. (Closed loop system I/O transfer functions and AI compensators). If $M_{ij} = 0$, $ij \neq 22$ and $M_{22} = D^{22}$ then $G_{11}(22) = P^{11}$, $G_{12}(22) = P^{21}$, $G_{22}(22) = G_{22}/G_{12}\ P^{21}$, and $G_{21}(22) = G_{21}/G_{12}\ P^{21}$, where $P^{ij} = -1/D^{ij}$.

Proposition 5. (Graphical evaluation of sensitivity reduction). The sensitivity operator matrix can be computed graphically by using the following expression

$$\$(M_{11}, M_{12}, M_{21}, M_{22}) = \$(M_{11})\ \$(M_{12})\Big|_{M_{11}}\ \$(M_{21})\Big|_{M_{11},M_{12}}\ \$(M_{22})\Big|_{M_{11},M_{12},M_{21}}$$

where

$$\$(M_{11}) = \begin{bmatrix} \dfrac{1/M_{11}}{1/M_{11}+G_{11}} & 0 \\ 0 & \dfrac{1/M_{11}+P^{11}}{1/M_{11}+G_{11}} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{1+G_{11}M_{11}} & 0 \\ 0 & \dfrac{M_{11} - D^{11}}{D^{11}(1+G_{11}M_{11})} \end{bmatrix}$$

$$\$(M_{12}) = \begin{bmatrix} \dfrac{1/M_{12}+P^{12}}{1/M_{12}+G_{21}} & 0 \\ 0 & \dfrac{1/M_{12}}{1/M_{12}+G_{21}} \end{bmatrix} = \begin{bmatrix} \dfrac{M_{12} - D^{12}}{D^{12}(1+G_{21}M_{12})} & 0 \\ 0 & \dfrac{1}{1+G_{21}M_{12}} \end{bmatrix}$$

$\$(M_{21})$, $\$(M_{22})$ are given by similar expressions, while $P^{ij} = -\dfrac{1}{D^{ij}}$.

Corollary. If $M_{12} = M_{21} = M_{22} = 0$ and M_{11} is such that in the frequency interval $[\omega_1, \omega_2]$

$$\left|\frac{\Delta_1}{\Delta_1^o}\right| < \alpha \text{ and } \left|\frac{\Delta_2}{\Delta_2^o}\right| < \beta,$$

then for $\omega \in [\omega_1, \omega_2]$ the Nyquist plot of $-1/M_{11}(j\omega)$ is interior to the circle with center at $\frac{-\alpha^2}{1-\alpha^2} G_{11}(j\omega)$ and radius $\frac{\alpha}{1-\alpha^2} |G_{11}(j\omega)|$ and to circle with center in $\frac{-\beta^2}{1-\beta^2} (G_{11}(j\omega) - P^{11}(j\omega))$ and radius $\frac{\beta}{1-\beta^2} |G_{11}(j\omega) - P^{11}(j\omega)|$.

An outline of the proof of all these propositions can be found in [5].

3. THE APPLICATIONS

Example ([5],[6]). The influence of small load variations over voltage and frequency deviations in a DC-AC converter of the type represented in Figure 5 can be studied in terms of the linear time invariant model represented in Figure 3, where

$$G_{11}(s) = \frac{\Delta F(s)}{\Delta U_1(s)} = \frac{.5125}{1+.55s} e^{-.08s}$$

$$G_{12}(s) = \frac{\Delta F(s)}{\Delta U_2(s)} = \frac{-.145}{1+.67s}\; c^{-.10s}$$

$$G_{21}(s) = \frac{\Delta V_a(s)}{\Delta U_1(s)} = \frac{1.775}{1+.41s}\; e^{-.04s}$$

$$G_{22}(s) = \frac{\Delta V_a(s)}{\Delta U_2(s)} = \frac{.675}{1+.32s} + \frac{.30s}{(1+.18s)^2}$$

$$F_1(s) = \frac{\Delta F(s)}{\Delta W(s)} = \frac{-.188}{1+.6s}\; e^{-.06s}$$

$$F_2(s) = \frac{\Delta V_a(s)}{\Delta W(s)} = \frac{-1.20}{1+.34s}$$

In this model η corresponds to the load variation ; $\overset{o}{\Delta}_1$ and $\overset{o}{\Delta}_2$ represent respectively the frequency and voltage deviations; U_1 denotes the voltage applied to the DC motor while U_2 is the voltage applied to the alternator excitation circuit. The effect of introducing a feedback compensator can be analyzed in terms of the scheme of Figure 4, where M_{11} and M_{22} can be respectively viewed as the frequency and voltage primary loops compensators, while M_{12} and M_{21} can be interpreted as the voltage-frequency and frequency-voltage intercoupling compensators.

Suppose the load variation η to be of the sinusoidal type with frequency $\omega = 2$ rad/sec: what is the performance improvement obtained by the introduction of a proportional compensator $M_{11} = 10$, while letting $M_{22} = M_{12} = M_{21} = 0$?

Proposition 5 suggests that this question can be answered graphically by using the Nyquist plots of G_{11}, $P^{11} = \frac{-(G_{21}F_1 - F_2G_{11})}{F_2}$ and $-1/M_{11}$. From Figure 6 we have

$$\left|\frac{\Delta_1}{\overset{o}{\Delta}_1}\right| = \frac{m_2}{m_1} = .24$$

$$\left|\frac{\Delta_2}{\overset{o}{\Delta}_2}\right| = \frac{m_3}{m_1} = .63.$$

This implies that a proportional frequency primary loop with $M_{11} = 10$ reduces the influence of a load variation of the type described above to 24% for the frequency deviation and to 63% for the voltage deviation.

Let us now assume a load variation of the sinusoidal type with frequency $\omega \in [0,10]$. Can we determine a univariate feedback compensator such that

$$\left|\frac{\Delta_1}{\overset{o}{\Delta}_1}\right| < .1 \quad \text{for} \quad 0 \leq \omega \leq 2$$

$$\left|\frac{\Delta_1}{\overset{o}{\Delta}_1}\right| < .3 \quad \text{for} \quad 2 \leq \omega \leq 10$$

and

$$\left|\frac{\Delta_2}{\overset{o}{\Delta}_2}\right| < .5 \quad \text{for} \quad 0 \leq \omega \leq 2$$

$$\left|\frac{\Delta_2}{\overset{o}{\Delta}_2}\right| < .65 \quad \text{for} \quad 2 \leq \omega \leq 10.$$

Proposition 5 suggests that this question can be answered via graphic considerations by using the Nyquist plot pairs: G_{11}, P^{11}; G_{21}, P^{12}; G_{12}, P^{21}; G_{22}, P^{22} where

$$P^{11} = \frac{-1}{D^{11}} = \frac{.5125\ e^{-.08s}}{1 + .55s} - \frac{1.775(.188)(1+.34s)e^{-.10s}}{1.20(1+.41s)(1+.6s)}$$

$$P^{12} = \frac{-1}{D^{12}} = \frac{1.775\ e^{-.04s}}{1 + .41s} - \frac{3.27\ (1+.6s)e^{-.02s}}{(1+.55s)(1+.34s)}$$

$$P^{21} = \frac{-1}{D^{21}} = \frac{-.145\ e^{-.10s}}{1 + .67s} - \frac{.675}{(1+.32s)} + \frac{.30s}{(1+.18s)^2}\ \frac{.188(1+.34s)e^{-.06s}}{1.20(1+.6s)}$$

$$P^{22} = \frac{-1}{D^{22}} = \frac{.675}{(1+.32s)} + \frac{.30s}{(1+.18s)^2} + \frac{.145(1.20)(1+.6s)e^{-04s}}{.188(1+.34s)(1+.67s)}$$

From these Nyquist plots pairs (see Figure 7) one can easily see that the answer is negative if one chooses the univariate compensator to be either M_{11}, M_{21} or M_{22}; it is positive if one chooses it to be M_{12}. After some simple analytical and graphical consideration one would for example find that the desired requirements are satisfied by choosing the lead-lag compensator

$$M_{12} = \frac{.7(1 + .3s)}{(1 + .03s)}.$$

Finally, let the load variation be once again of the sinusoidal type with frequency $\omega = 2$ rad/sec. Suppose that we have $M_{12} = 2$, $M_{22} = 1.25$, $M_{11} = M_{21} = 0$: is the performance of the compensated system better than that of the original open loop system? This question can once again be answered graphically by using Proposition 5. To do this it is sufficient to draw the Nyquist plots of G_{21}, P^{12} and $-1/M_{12}$ plus the Nyquist plots of

$$G_{22}(12) = \frac{G_{22}}{1+G_{21}M_{12}},\quad P^{22}(12) = \frac{P^{22}}{1+M_{12}P^{12}},\ \text{and}\ -1/M_{22}.$$

Following then the same procedure illustrated above from these Nyquist plots we obtain (Figures 7 and 8)

$$\left|\frac{\Delta_1}{\overset{o}{\Delta}_1}\right| = .6 \times .3 = .18$$

$$\left|\frac{\Delta_2}{\Delta_2^o}\right| = .28 \times .8 = .23.$$

Clearly the compensated system performance is considerably better than that of the original system. A full development of this example can be found in [5]; a detailed description of the model used is given in [6].

4. CONCLUSIONS

As it was to be expected from their definitions and their subsequent expressions and properties, the Absolute Invariant, (AI), compensators are not always well defined. When they are well defined they are not necessarily physically realizable, nor do they in general make the closed loop system stable. In addition to this, the choice of some specific combinations of AI compensators leads inequivocally to instability. The surprising aspect of the present development is that in spite of all these difficulties, the analysis of the AI compensators can indeed play a useful role in feedback system design. On the one hand, when the AI compensators are causal and well defined, it may occasionally be advantageous to insert one or two of them in the feedback loop. On the other hand, it is useful to consider the AI compensators even when we have no intention whatsoever of inserting them in the feedback loop: this independently from their stability or causality properties.

The fact is that the effect of a general compensator M_{ij} over the behavior of the system can be conveniently discussed either graphically or analytically in terms of its relative distance from D^{ij} and of the gain of the (additional) feedback loop which is created by the presence of M_{ij}. This is particularly true if one is interested in the following type of comparative sensitivity questions:

- Given G_{ij}, F_i, M_{ij} is the compensated system better than the open loop system?
- Given that we want to use only one compensator, which one of the four should we pick and how?
- Given three of the four feedback compensators, how should one choose the forth one?

5. REFERENCES

[1] Some Considerations Relating to Control Systems Employing the Invariance Principle, Preminger, J., Rootenberg, J., IEEE Tr. on AC, July 1964.

[2] Benchmark Papers in Systems Sensitivity Analysis, J.B. Cruz, (Editor), Dowdon, Hutchison & Ross, Inc., 1974.

[3] The Return Difference and Return Ratio Matrices and Their Use in the Analysis and Design of Multivariable Control Systems, Rosenbrock,H.H., Proc. IEEE 117, 2027-2049, 1970.

[4] Synthesis of Multivariable, Basically Non-Interacting Synthesis with Significant Plant Uncertainty, Shaked, U., Horowitz, I., Goldes, S., Automatica, Vol. 12, pp. 61-71, 1976.

[5] Techniques Fréquentielles pour l'Analyse des Systèmes Multivariables Continus avec Application au Réglage Fréquence-Tension dans un Groupe Moteur-Alternateur, De Santis, R.M., Lefebvre, S., TR Ecole Polytechnique de Montréal, Septembre 1976.

[6] Commande par Ordinateur d'un Groupe Moteur-Alternateur, Lefebvre, S., Projet de Fin d'Etude, Ecole Polytechnique de Montréal, April 1976.

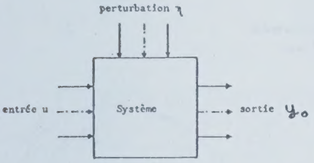

perturbation η
entrée u
Système
sortie y_o

Figure 1: Open loop system

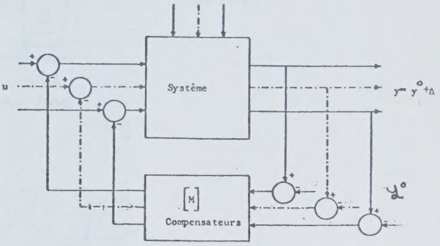

u
Système
$y = y^o + \Delta$
[M]
Compensateurs
y^o

Figure 2: Closed loop system

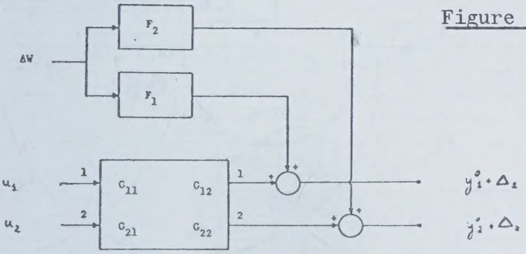

ΔW
F_2
F_1
u_1
u_2
1
2
G_{11}
G_{12}
G_{21}
G_{22}
1
2
$y_1^o + \Delta_1$
$y_2^o + \Delta_2$

Figure 3: 2 Inputs - 2 Outputs Open loop system

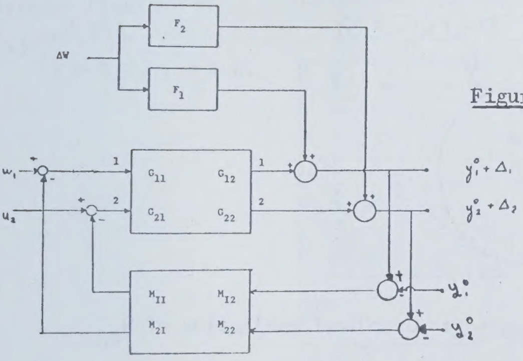

ΔW
F_2
F_1
w_1
u_2
1
2
G_{11}
G_{12}
G_{21}
G_{22}
1
2
$y_1^o + \Delta_1$
$y_2^o + \Delta_2$
M_{11}
M_{12}
M_{21}
M_{22}
y_1^o
y_2^o

Figure 4: 2 Inputs-2 Outputs Closed loop system

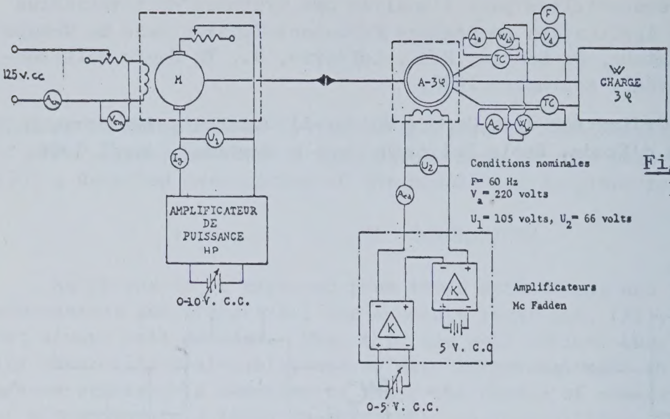
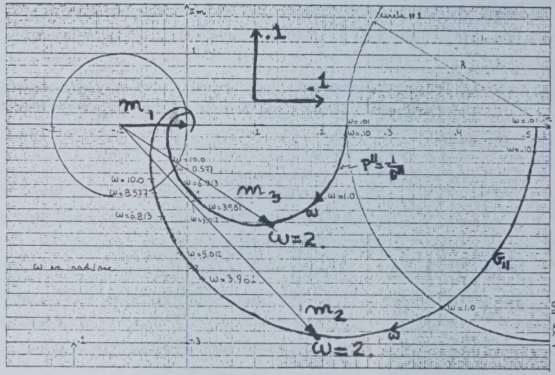

125 v. cc
M
A-3φ
CHARGE 3φ
F
TC
TC
AMPLIFICATEUR DE PUISSANCE HP
0-10 V. C.C.
Conditions nominales
F= 60 Hz
V_a= 220 volts
U_1= 105 volts, U_2= 66 volts
Amplificateurs
Mc Fadden
K
K
5 V. C.C
0-5 V. C.C.

Figure 5: Open loop DC-AC Converter

Figure 6: Graphical evaluation of M_{11}



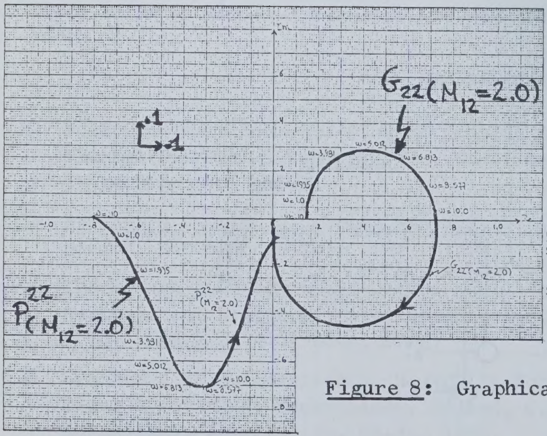
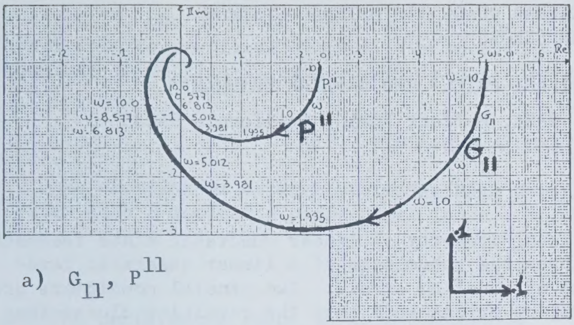
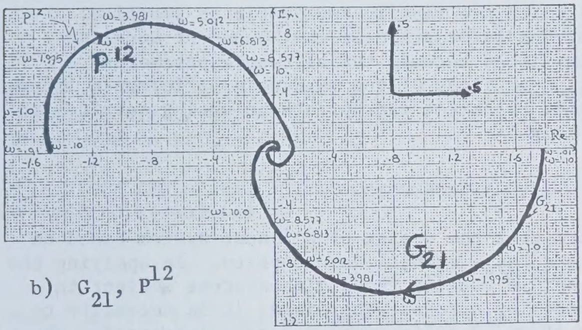


Figure 8: Graphical evaluation of M_{22}



a) G_{11}, P^{11}



b) G_{21}, P^{12}

c) G_{12}, P^{21}

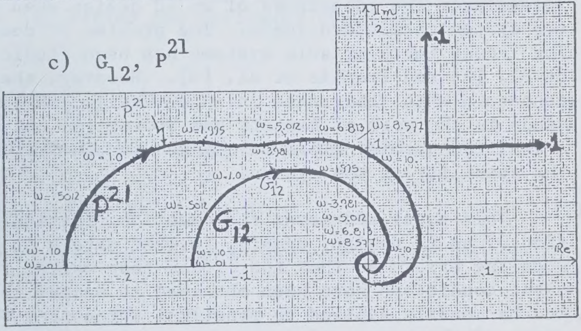
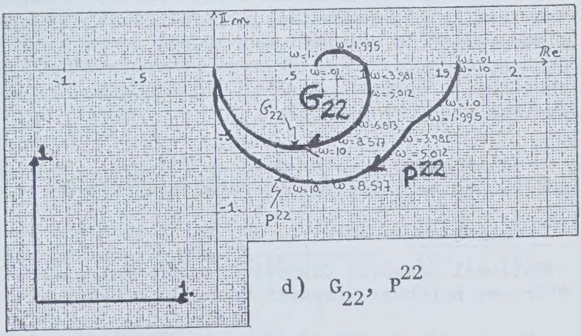


Figure 7:

Nyquist Diagrams of G_{ij} and P^{ij}, $i = 1,2$; $j = 1,2$.



d) G_{22}, P^{22}

CLOSED-LOOP STABILITY OF LINEAR QUADRATIC OPTIMAL SYSTEMS IN THE PRESENCE OF MODELING ERRORS

MITSUHIKO TODA,* RAJNIKANT PATEL* AND BANAVAR SRIDHAR**

Ames Research Center, NASA, Moffett Field, California 94035

ABSTRACT

The well-known stabilizing property of linear quadratic state feedback design is utilized to evaluate the robustness of a linear quadratic feedback design in the presence of modeling errors. Two general conditions are obtained for allowable modeling errors such that the resulting closed-loop system remains stable. One of these conditions is applied to obtain two more particular conditions which are readily applicable to practical situations where a designer has information on the bounds of modeling errors. Relations are established between the allowable parameter uncertainty and the weighting matrices of the quadratic performance index, thereby enabling the designer to select appropriate weighting matrices to attain a robust feedback design.

I. INTRODUCTION

It is widely recognized that the design of a linear system based on linear-quadratic (LQ) methods has a stabilizing property. In applying the LQ design, a control designer is usually faced with systems subject to parameter variations and modeling errors. Therefore, it is necessary to investigate the robustness of an LQ design when accurate information about the system is not available. The problem of designing robust controllers for linear multivariable systems has been studied by Davison [1], Staats et al. [2] and Francis et al. [3]. However, their work is not specifically concerned with robustness in an LQ design. More relevant to the present study is the work of Barnett and Storey [4] and Wong and Athans [5] who have classified special kinds of changes in the system matrix which do not affect the stability of the system.

This paper addresses the question, inevitable in the practical application of an LQ design, namely, to what extent an LQ design can tolerate modeling errors so that the closed-loop system remains stable. Four sufficient conditions have been obtained to answer this question. Two of them can be readily applied to practical situations where a designer has estimates of the bounds of system parameter perturbations. Relations have been established between the allowable parameter uncertainty and the weighting matrices in the quadratic performance index. As a consequence, the conditions will help the designer select appropriate weighting matrices in the quadratic performance index to attain a robust feedback design. We believe that these results are useful in the computer aided design of multivariable systems using an LQ design.

This paper is organized as follows: Section II formulates the problem, and two general sufficient conditions are obtained in section III. In

---

*National Research Council Research Associate, Ames Research Center.
**Former National Research Council Research Associate, Ames Research Center.

section IV, one of these conditions is used to obtain two conditions readily applicable to practical designs. In section V, we present two examples and conclude in section VI with remarks about this design procedure.

II. PROBLEM STATEMENT

We consider an optimal state feedback design for a linear time invariant system with a quadratic performance index. The system equation is

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{1}$$

and the performance index to be minimized is

$$J = \int_0^\infty [x(t)^T Qx(t) + u(t)^T Ru(t)]dt \tag{2}$$

where $x(t) \in R^n$, $u(t) \in R^m$, and the weighting matrices Q and R are assumed positive definite (denoted $Q > 0$ and $R > 0$).* Further, the system is assumed to be controllable.

Since the system parameters are usually unknown, the system (1) is assumed to be modeled by

$$\dot{x}(t) = A_m x(t) + Bu(t) \tag{3}$$

where $A_m \equiv A + E$ and E denotes the modeling error. The optimization of index (2) with system model (3) yields the following constant gain state feedback control

$$u(t) = -R^{-1}B^T Px(t) \tag{4}$$

where P is the solution of the following algebraic Riccati equation

$$A_m^T P + PA_m - PBR^{-1}B^T P + Q = 0 \tag{5}$$

It is well known that the resulting closed-loop system with control (4)

$$\dot{x}(t) = (A - BR^{-1}B^T P)x(t) \tag{6}$$

is stable if $E = 0$, namely when there is no modeling error. The system model (3), however, is usually subject to errors represented by E. Therefore, it is important to know how robust the design is in the presence of modeling errors; i.e., how much parameter variations E can be tolerated so that the closed-loop system (6) remains stable? This question is answered in the following two sections.

III. REGIONS OF MODELING ERROR FOR CLOSED-LOOP STABILITY

In this section we obtain two regions of the modeling error where stability of the resulting closed-loop system is guaranteed. One of them will be specialized in the next section to generate a region which can be readily used for practical designs.

---

*Similarly, a positive semidefinite matrix W will be denoted by $W \geq 0$.

One of the stability regions (Region 1) is specified by

Theorem 1

If the perturbation E satisfies the condition

$$\|E\|_s \leq \frac{1}{2\|P\|_s\|T^{-1}\|_s} = \frac{\operatorname{Min} \lambda(T)}{2 \operatorname{Max} \lambda(P)} \tag{7}$$

then the closed-loop system (6) is stable, where $\|W\|_s$ and $\lambda(W)$ denote the spectral norm and an eigenvalue of a matrix W, and $T \equiv 2Q + A_m^T P + PA_m$.

Remarks:

1. The spectral norm $\|W\|_s$ is the maximum singular value of W, denoted by $\operatorname{Max} \alpha(W)$ [6, p. 57];

$$\|W\|_s = \operatorname{Max} \alpha(W) \equiv \operatorname{Max}[\lambda(WW^T)]^{1/2}$$

and when W is normal [7, p. 46]

$$\|W\|_s = \rho(W) = \operatorname{Max}|\lambda(W)| \text{ (Spectral radius of } W)$$

Furthermore, when W is positive definite

$$\|W^{-1}\|_s = \operatorname{Max} \lambda(W^{-1}) = [\operatorname{Min} \lambda(W)]^{-1}$$

2. From the Riccati equation (5) $T = Q + PBR^{-1}B^TP$. Since $Q > 0$, $T > 0$, the condition (7) bounds $\|E\|_s$ by some positive number. Also since $T - Q \geq 0$, $\operatorname{Min} \lambda(T) \geq \operatorname{Min} \lambda(Q) > 0$. Therefore, $\|T^{-1}\|_s$ can be replaced by $\|Q^{-1}\|_s$ in (7) and in the bounds in the following theorems to obtain sufficient conditions for stability.

Proof of Theorem 1

Since $A_m = A + E$ the Riccati equation (5) can be rewritten as

$$\begin{aligned}(A - BR^{-1}B^TP)^TP + P(A - BR^{-1}B^TP) &= -(Q + PBR^{-1}B^TP + E^TP + PE) \\ &= -(T + E^TP + PE) \qquad (8)\end{aligned}$$

Since P is positive definite,

$$T + E^TP + PE \geq 0 \tag{9}$$

is sufficient for (6) to be stable, i.e., $\operatorname{Re}\lambda(A - BR^{-1}B^TP) \leq 0$, [4, p. 77].

Applying the minimax property of eigenvalues to the symmetric matrices T and $E^TP + PE$ [6, p.102],* a sufficient condition for (9) is

$$\text{Min } \lambda(T) \geq \text{Max}|\lambda(E^TP + PE)|$$

The spectral radius of $E^TP + PE$ is bounded by

$$\begin{aligned} &\text{Max}|\lambda(E^TP + PE)| \\ &\quad \leq [\text{Max } \lambda(PEE^TP)]^{1/2} + [\text{Max } \lambda(E^TPPE)]^{1/2} = 2[\text{Max } \lambda(PEE^TP)]^{1/2} \\ &\quad \leq 2[\text{Max } \lambda(EE^T)\text{Max } \lambda(P^2)]^{1/2} \\ &\quad = 2 \text{ Max } \alpha(E)\text{Max } \lambda(P) \end{aligned}$$

where the two inequalities are due to Wittmeyer [8, pp. 66, 67], and the last equality is obtained from the positive definiteness of P.

Thus,

$$\text{Min } \lambda(T) \geq 2 \text{ Max } \alpha(E)\text{Max } \lambda(P)$$

suffices for the stability of (6), which completes the proof.

The other sufficient condition for stability (Region 2) is stated in the following theorem, which is proved in the APPENDIX.

Theorem 2

If the perturbation E satisfies the condition

$$\frac{\text{Max } \lambda[(I + E^T)(I + E)]}{\text{Min } \lambda[(I - E^T)(I - E)]} \leq \rho_1 + \frac{2\rho_2}{\text{Max } \lambda[(I - E^T)(I - E)]} \tag{10}$$

where I is the identity matrix and

$$\rho_1 \equiv [\|P^{-1}\|_s\|P\|_s]^{-1}, \quad \rho_2 \equiv [\|T^{-1}\|_s\|P\|_s]^{-1},$$

then the closed-loop system is stable.

Applying the minimax property to the symmetric matrices $E^T + E$ and E^TE, (10) yields another sufficient condition:

Corollary 1

If the perturbation E satisfies the condition

---

*If n x n matrices B and C are symmetric, then

$$\lambda_i(B) + \text{Min } \lambda(C) \leq \lambda_i(B + C) \leq \lambda_i(B) + \text{Max } \lambda(C), \quad i = 1, \ldots, n,$$

where $\lambda_i(B)$ and $\lambda_i(B + C)$ are the ith eigenvalues of B and B + C, respectively, when the eigenvalues are arranged in increasing or decreasing order.

$$\frac{1 + \text{Max}\ \lambda(E^T + E) + \text{Max}\ \lambda(E^TE)}{1 - \text{Max}\ \lambda(E^T + E) + \text{Min}\ \lambda(E^TE)} \le \rho_1 + \frac{2\rho_2}{1 - \text{Min}\ \lambda(E^T + E) + \text{Max}\ \lambda(E^TE)}$$

then the closed-loop system is stable.

Remarks:

1. Theorem 1 is more useful, since it clearly contains the null matrix ($E = 0$), whereas Theorem 2 does not necessarily include the null matrix. (The inequality (10) requires $1 \le \rho_1 + 2\rho_2$ if $E = 0$; see Example 2 of section V.)

2. Region 2 may not be bounded in the region of stability matrices, while Region 1 is always bounded (Example 1 of section V).

3. Note that nonsingularity of $I - E$ is required for (10) to be meaningful ($\text{Re}\ \lambda(E) < 1$ is sufficient).

IV. BOUNDS ON MODELING ERROR FOR PRACTICAL DESIGN

In applying the linear quadratic feedback design technique to dynamic systems, a designer usually has information on the bounds on each element of the modeling error matrix E. In this section we obtain conditions which the designer can readily apply to evaluate the robustness of his design for such situations. The following results require knowledge of the bounds on elements of E and do not require any eigenvalue computations for E.

Theorem 3 (Region 3)

If the perturbation E satisfies the condition

$$\|E\|_E \le \frac{1}{2\|P\|_s\|T^{-1}\|_s} \tag{11}$$

where $\|E\|_E = \left(\sum_{i=1}^{n}\sum_{j=1}^{n} |e_{ij}|^2\right)^{1/2}$ is the Euclidean norm of the matrix E whose (i,j) element is e_{ij}, then the closed-loop system is stable.

Theorem 3 is clear, since $\|E\|_s \le \|E\|_E$ in (7) [6, p. 57]. The result contains the following important class of perturbations.

Theorem 4 (Region 4)

If the perturbation E satisfies the condition

$$|e_{ij}| \le \varepsilon = \frac{1}{2n} \cdot \frac{1}{\|P\|_s\|T^{-1}\|_s} \qquad \forall i,j = 1, \ldots, n,$$

then the closed-loop system (6) is stable.

Proof of Theorem 4

The result follows directly from the definition of $\|E\|_E$.

Remarks:

1. In [1], Davison specifies a class of perturbation of the form $\left\{e_{ij} \middle| |e_{ij}| < \varepsilon\right\}$ to define robust controllers.

2. Note that the bound ε is inversely proportional to n, the dimension of the state $x(t)$.

V. EXAMPLES

Example 1

For a scalar system Region 1 reduces to

$$-\left(A_m + \frac{Q}{P}\right) \leq E \leq A_m + \frac{Q}{P}$$

Region 2 reduces to

$$E \leq A_m + \frac{Q}{P} \quad \left(A_m + \frac{Q}{P} > 0\right)$$

which is the necessary and sufficient condition obtained by direct evaluation.

Example 2

Consider the following 2-dimensional system with single input

$$A_m = \begin{bmatrix} 0 & 1 \\ 1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad Q = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}, \quad R = 1$$

For this system

$$\rho_1 = 0.19 \ , \quad \rho_2 = 0.32$$

Therefore, the regions of stability are specified by

Region 1: $\quad \text{Max } \alpha(E) \leq \frac{1}{2}\rho_2 = 0.16$

Region 2: $\quad \dfrac{\text{Max } \lambda[(I + E^T)(I + E)]}{\text{Min } \lambda[(I - E^T)(I - E)]} \leq 0.19 + \dfrac{0.64}{\text{Min } \lambda[(I - E^T)(I - E)]}$

Region 3: $\quad \left(e_{11}^2 + e_{12}^2 + e_{21}^2 + e_{22}^2\right)^{1/2} \leq 0.16$

Region 4: $\quad |e_{ij}| \leq \varepsilon = \frac{1}{4}\rho_2 = 0.08$

Note that Region 2 does not include the null matrix, since $1 \leq \rho_1 + 2\rho_2$ is not satisfied. The least upper bound ε for the class of perturbations $\left\{e_{ij} \middle| |e_{ij}| \leq \varepsilon\right\}$ corresponding to Region 4 is 0.2, which can be obtained by direct evaluation of the eigenvalues for this example.

Writing the state weighting matrix Q as qI, where q is a positive number, we can show that $d\rho_2/dq > 0$ and $\lim_{q \to +\infty} (1/2)\rho_2 = 0.5$; thus it follows that more penalty on the states results in more robustness to modeling errors and that there exists an upper bound for ρ_2. In general, ρ_2 of an LQ design is bounded by a function of the dominant pole of the closed-loop matrix $A_m - BR^{-1}B^TP$ [9].

VI. CONCLUDING REMARKS

By making use of the stabilizing property of the LQ design, four sufficient conditions were obtained for the stability of a closed-loop system in the presence of modeling errors. Barnett and Storey [4] and Wong and Athans [5] have considered a similar problem. However, their results are different from ours in that the changes which they allow in the system matrix to preserve closed-loop stability are of the special form $(N - M)K$ where $N = -N^T$, $M \geq 0$ and $K \geq 0$ satisfying $K A + A^TK \leq 0$. In a practical design problem it may be difficult to relate the matrices N, M and K to bounds on the error matrix E. The results which we have presented in this paper are not restricted to any special form of the error matrix and consequently are of wider practical applications.

Although the modeling error matrix E has been assumed constant in this paper, it is easy to see that the results obtained also hold for time varying E. The extension to nonlinear errors is reported in [9] together with an application of the results of this paper for the LQ design of autopilots for a STOL aircraft.

From the point of view of design, the robustness result in Theorem 4 is the most useful. The only computation required in obtaining a measure of the robustness of any LQ design is the determination of the eigenvalues of two symmetric matrices. The matrices themselves are obtained in the course of solving the algebraic Riccati equation corresponding to the linear quadratic design problem. Hence the result in Theorem 4 can be easily incorporated in any LQ design with little extra computation, and is therefore particularly suitable for use in a computer-aided design environment. The main limitation of the result is that it involves P, the solution of the algebraic Riccati equation, thereby making the result an implicit relation between the bound ε and the weighting matrices Q and R. This necessarily requires the solution of an algebraic Riccati equation in order to compute ε. In this context, research to express stability regions explicitly in terms of the system parameters (A and B) and weighting matrices (Q and R) is desired. It should also be pointed out that the results presented in this paper give only sufficient conditions for closed-loop stability in the presence of modeling errors, and these may often be very conservative with regard to the allowable bounds on the modeling error matrix E.

APPENDIX — PROOF OF THEOREM 2

Consider the Cayley transformation

$$C = (I - E)^{-1}(I + E)$$

for nonsingular $I - E$, then E is expressed by

$$E = (C - I)(C + I)^{-1}$$

Substituting this into the right-hand side of equation (8), writing $S \equiv -(T + E^TP + PE)$, and premultiplying by $(C^T + I)$ and postmultiplying by $(C + I)$, we obtain

$$(C^T + I)S(C + I) = -2\left[-C^TPC + P + \frac{1}{2}(C^T + I)T(C + I)\right]$$

A sufficient condition for closed-loop stability (inequality (9) in the proof of Theorem 1) is satisfied if the right-hand side of the above equation is negative semidefinite. Applying the minimax property of eigenvalues to the symmetric matrices on the right-hand side, the following condition results:

$$\text{Max } \lambda(C^TPC) \leq \text{Min } \lambda(P) + \frac{1}{2} \text{Min } \lambda[(C^T + I)T(C + I)] \tag{12}$$

Bounds on the eigenvalues of matrices involving C can be obtained by applying Wittmeyer's inequalities [8, p. 66]

$$\begin{aligned} \text{Max } \lambda(C^TPC) &\leq \text{Max } \lambda(P)\text{Max } \lambda(CC^T) \\ &\leq \text{Max } \lambda(P)\text{Max } \lambda[(I + E^T)(I + E)]\text{Max } \lambda[(I - E^T)^{-1}(I - E)^{-1}] \\ &= \text{Max } \lambda(P) \frac{\text{Max } \lambda[(I + E^T)(I + E)]}{\text{Min } \lambda[(I - E^T)(I - E)]} \end{aligned} \tag{13}$$

$$\begin{aligned} \text{Min } \lambda[(C^T + I)T(C + I)] &\geq \text{Min } \lambda[(C^T + I)(C + I)]\text{Min } \lambda(T) \\ &= \frac{4 \text{ Min } \lambda(T)}{\text{Max } \lambda[(I - E^T)(I - E)]} \end{aligned} \tag{14}$$

where the equality $C + I = 2(I - E)^{-1}$ was used to obtain the last equality.

The inequalities (12), (13) and (14) yield (10), completing the proof.

REFERENCES

1. Davison, E. J.: "The robust control of a servomechanism problem for linear time-invariant multivariable systems," IEEE Trans. Automat. Contr., Vol. AC-21, pp. 25-34, Feb. 1976.

2. Staats, Jr., P. W. and Pearson, J. B.: "Robust solution of the linear servomechanism problem," Dep. Elec. Eng., Rice Univ., Houston, Texas, Tech. Rep. 7401, Sept. 1974.

3. Francis, B., Sebakhy, O. A. and Wonham, W. M.: "Synthesis of multivariable regulators: The internal model principle," Appl. Math. Optimiz., Vol. 1, No. 1, pp. 64-86, 1974.

4. Barnett, S. and Storey, C.: "Matrix methods in stability theory," Barnes and Noble, New York, 1970.

5. Wong, P. K. and Athans, M.: "Closed-loop structural stability for linear-quadratic optimal systems," submitted to 1976 IEEE Conf. Decision and Control.

6. Wilkinson, J. H.: "The algebraic eigenvalue problem," Oxford Univ. Press, London, 1965.

7. Householder, A. S.: "The theory of matrices in numerical analysis," Dover, New York, 1975.

8. Bodewig, E.: "Matrix Calculus," North-Holland, Amsterdam, 1956.

9. Patel, R., Sridhar, B. and Toda, M.: "Robustness in linear quadratic feedback design with application to an aircraft control problem," submitted to the Tenth Asilomar Conference on Circuits, Systems and Computers, 1976.

ON THE ROBUST CONTROL OF A TIME DELAY SERVOMECHANISM PROBLEM[†]

A. GOLDENBERG** and E.J. DAVISON*

*Dept. of Electrical Engineering
University of Toronto
Toronto, Canada M5S 1A4
**Spar Aerospace Ltd., Toronto, Canada

ABSTRACT

In this paper, the servomechanism problem for a linear time-invariant multivariable system is extended to systems with time delay. Necessary and sufficient conditions for the existence of a solution to the time delay servomechanism problem are given and the structure of compensators required for this problem is obtained. The robust time delay servomechanism problem is then studied, and a characterization of all robust controllers is then made.

1. INTRODUCTION

In this paper, the servomechanism problem defined in [1]-[3], is extended to systems with delay - the time delay servomechanism problem. Necessary and sufficient conditions for the existence of a solution to the time delay servomechanism problem are given. The robust time delay servomechanism problem is also solved and a characterization of robust compensators is made.

The following linear, time invariant system is considered

$$\begin{aligned} \dot{x} &= A_0 x + A_1 x(t-\Delta) + Bu + Ev \\ e &= C_0 x + C_1 x(t-\Delta) + Du + Fv \\ y_m &= C_{0_m} x + C_{1_m} x(t-\Delta) + F_m v \end{aligned} \tag{1.1}$$

where $x \in \mathbb{R}^n$ is the state-trajectory, $u \in \mathbb{R}^m$ is the control input, $e \in \mathbb{R}^r$ is the error, $y_m \in \mathbb{R}^{r_m}$ is the measurable output and $v \in \mathbb{R}^\ell$ is the external input given by:

$$\dot{v} = Hv \tag{1.2}$$

where $H \in \mathbb{R}^{\ell\times\ell}$ and where for all $\lambda \in \sigma(H)$, $\text{Re}\,\lambda \geq 0$, where $\sigma(\cdot)$ denotes the spectrum of $(\cdot)$. The initial condition of (1.1) is a continuous function $\phi(\theta)$ defined over the interval $[-\Delta,0]$. This function defines the state space $C^n \triangleq C([-\Delta, 0]; \mathbb{R}^n)$, the n vector valued continuous function over the interval $[-\Delta,0]$ with norm $\|\phi\| = \max_{\theta\in[-\Delta,0]}|\phi(\theta)|$. The state of the system (1.1) is then defined as the segment $x(t+\theta)$, $\theta\in[-\Delta,0]$, of the state trajectory x. In the development to follow, the state will be denoted as $x_t(\theta) \triangleq x(t+\theta)$ for all $t \in [0,\infty)$. In particular, the initial condition is denoted by x_0 which implies that $x_0 = \phi(\theta)$, $\theta\in[-\Delta,0]$. In [11], the space $M_2 \triangleq \mathbb{R}^n \times L_2([-\Delta, 0], \mathbb{R}^n)$ is suggested as the most suitable set up for analyzing time delay systems.

[†] This work has been supported by the National Research Council of Canada under Grant No. A4396.

2. PRELIMINARIES

The basic mathematical set up considered is the linear, time invariant, functional differential equation (FDE) described as follows [4]:

$$x(t) = L(x_t) + f(t) \tag{2.1}$$

where $L: C^n \to \mathbb{R}^n$ is a linear bounded operator and $f \in L_1^{loc}([0,\infty); \mathbb{R}^n)$. By Riesz's representation theorem [5]:

$$\dot{x}(t) = \int_{-\Delta}^{0} d\eta(\theta) x(t+\theta) + f(t) \tag{2.2}$$

where $\eta(\theta)$ is a matrix valued function of bounded variation. In particular, if

$$[d\eta(\theta)]\,(\cdot) \triangleq [A_0\delta(\theta) + A_1\delta(\theta+\Delta)]\,(\cdot)d\theta$$

where $\delta(\theta)$ is the unit impulse function, and if

$$f(t) \triangleq Bu(t)$$

then (2.2) becomes the time delay system as defined by (1.1)

$$\dot{x} = A_0 x + A_1 x(t-\Delta) + Bu \tag{2.3}$$

where $\qquad u \in \mathcal{U} \triangleq L_1^{loc}([0,\infty); \mathbb{R}^m)$

The technique which uses $x_t \in C^n$ as a state space rather than $x \in \mathbb{R}^n$, for time delay systems, has been introduced in [6] and later used in [4]. The basic difference is that the state x_t represents a segment of trajectory x. Systems described by (1.1) can therefore be analyzed in their natural set up $\mathbb{R}^n$.

Let $x(t_0,\phi,f)$ be the solution of (2.1) with initial value ϕ at t_0. For convenience, we shall let $t_0 \equiv 0$. Since the operator L is linear and by the uniqueness of the solution of (2.1) [4], it follows that:

$$x(\phi,f) = x(\phi,0) + x(0,f)$$

If $f \in \mathcal{U}$ and $x(\phi,f)$ is the solution of (2.1) then [4]

$$x(\phi,f)(t) = x(\phi,0)(t) + \int_0^t Q(t-s)f(s)ds \tag{2.4}$$

$$x_0 = \phi$$

where Q(t), the solution of the fundamental equation of (2.1) with $f(t) \equiv 0$, is given as follows:

$$\dot{Q}(t) = L(Q_t) \qquad \text{for } t > 0$$

$$Q(t) = I \qquad \text{for } t = 0$$

$$Q(t) = 0 \qquad \text{for } t < 0$$

where $\qquad Q_t(0) \triangleq Q(t+\theta),\ \theta \in [-\Delta,0].$

It is convenient to write (2.4) as follows:

$$x(\phi,f)(t+\theta) = x(\phi,0)(t+\theta) + \int_0^{t+\theta} Q(t+\theta-s)f(s)ds$$

$$x(\phi,f)(t+\theta) = \phi(t+\theta) \qquad\qquad t+\theta\in[-\Delta,0]$$

or, according to the definition of the state space C^n

$$x_t(\phi,f)(\phi) = x_t(\phi,0)(\theta) + \int_0^t Q_t(\cdot,s)(\theta)f(s)ds$$

or, using a more general notation:

$$x_t(\phi,f) = x_t(\phi,0) + \int_0^t Q_t(\cdot,s)f(s)ds$$

Since $\phi \in C^n$ and $x_t(\phi,0)$ is the unique solution of (2.1) with $f(t) \equiv 0$, one may define the operator $T:C^n \rightarrow C^n$ as follows

$$x_t(\phi) = T(t)\phi \qquad\qquad t \geq 0 \tag{2.5}$$

Stabilizability for time delay systems is defined as follows [7]:

The system (2.3) is stabilizable if there exists a bounded operator K, K: $C^n \rightarrow U$, such that $u(t) = -Kx_t$ makes the system asymptotically stable, i.e. $\lim_{t\to\infty} x_t = 0$ for all initial conditions $x_0 \in C^n$.

A necessary and sufficient condition for stabilizability of time delay systems (1.1) is given in the following statement [8], [9]:

The time delay system (1.1) is stabilizable iff for all $\lambda \in \sigma(A_0+A_1e^{-\lambda\Delta})$ with $\text{Re}\ \lambda \geq 0$ the following condition holds:

$$\text{rank}\ (A_0+A_1e^{-\lambda\Delta}-\lambda I,\ B) = n \tag{2.6}$$

Observability of time delay systems is now defined [10]. Let $P_0:C^n \rightarrow \mathbb{R}^n$ so that $P_0x_t(\theta) \triangleq x_t(0)$, $\theta\in[-\Delta,0]$ and $P_1:C^n \rightarrow \mathbb{R}^n$ so that $P_1x_t(\theta) \triangleq x_t(-\Delta)$. Also let $\bar{C}_0 \triangleq C_0P_0$ and $\bar{C}_1 = C_1P_1$. Then the output equation in (1.1) is equivalent to

$$y = (\bar{C}_0 + \bar{C}_1)x_t \triangleq \bar{C}x_t \tag{2.7}$$

The system (2.3), (2.7) is observable if for all $x_0 \in C^n$

$$\bar{C}\ T(t)x_0 = 0 \qquad\qquad t \geq 0 \tag{2.8}$$

implies $x_0 = 0 \in C^n$.

A necessary and sufficient condition for detectability of time delay systems (1.1) is given in the following statement [8], [9]:

The system (2.3), (2.7) is detectable iff for all $\lambda \in \sigma(A_0+A_1e^{-\lambda\Delta})$ with $\mathrm{Re}\,\lambda \geq 0$ the following holds*:

$$\mathrm{rank}\begin{bmatrix} A_0+A_1e^{-\lambda\Delta}-\lambda I \\ C_0+C_1e^{-\lambda\Delta} \end{bmatrix} = n \tag{2.9}$$

Observers for time delay systems have been extensively treated in [10], [12]. These results will be used in the development to follow.

3. THE SOLVABILITY OF THE TIME DELAY SERVOMECHANISM PROBLEM

In this section the time delay servomechanism problem is defined and solved.

In order to define the time delay servomechanism problem we make the following definition of an extended system combining (1.1) with the external input (1.2)

$$\dot{x}_e = \begin{pmatrix} A_0 & E \\ 0 & H \end{pmatrix} x_e + \begin{pmatrix} A_1 & 0 \\ 0 & 0 \end{pmatrix} x_e(t-\Delta) + \begin{pmatrix} B \\ 0 \end{pmatrix} u \tag{3.1}$$

$$e = (C_0 \quad F)x_e + (C_1 \quad 0)x_e(t-\Delta) + Du$$

$$y_m = (C_{0_m} \quad F_m)x_e + (C_{1_m} \quad 0)x_e(t-\Delta) + D_m u$$

The following servomechanism problem is now defined. Given the time delay, time invariant, linear system described by (3.1), find necessary and sufficient conditions for the existence of a state feedback $K^e : \mathcal{C}^{n+\ell} \to \mathcal{U}$ defined as follows

$$u = K_0 x + \int_{-\Delta}^{0} K_1(\theta)x(t+\theta)d\theta + K_2 v \tag{3.2}$$

so that the following hold:

a) the system is closed loop asymptotically stable

b) output regulation occurs, i.e.

$$\lim_{t\to\infty} e(t) = 0 \qquad \forall x_t^e(0) \in \mathcal{C}^{n+\ell}$$

The problem defined above may be referred to as the time delay feedforward servomechanism problem with complete state measurement.

Consider now the closed loop time delay system obtained by combining (3.1) with (3.2). The following is obtained:

$$\dot{x}_e = \begin{pmatrix} A_0+BK_0 & E+BK_2 \\ 0 & H \end{pmatrix} x_e + \begin{pmatrix} A_1 & 0 \\ 0 & 0 \end{pmatrix} x_e(t-\Delta) + \int_{-\Delta}^{0} \begin{pmatrix} B \\ 0 \end{pmatrix} K_1(\theta)x(t+\theta)d\theta$$

$$e = (C_0+DK_0 \quad F+DK_2)x_e + (C_1 \quad 0)x_e(t-\Delta) + \int_{-\Delta}^{0} DK_1(\theta)x(t+\theta)d\theta \tag{3.3}$$

---

* Due to space limitations, all proofs are omitted; they may be found in [8].

Using now the formulation (3.3), after Laplace transformation, the requirement a) becomes

$$\sigma\Big[A_0+BK_0+A_1e^{-\lambda\Delta}+\int_{-\Delta}^{0}K_1(\theta)e^{\lambda\theta}d\theta\Big]\subset\mathbb{C}^- \qquad (3.4)$$

Applying now the Laplace transform to (3.3) the following transfer function $G_c(s)$ of the closed loop time delay system is obtained:

$$G_c(s)\triangleq\Big[C_0+DK_0+C_1e^{-s\Delta}+\int_{-\Delta}^{0}DK_1(\theta)e^{s\theta}\ \theta\Big]\cdot$$

$$\Big[A_0+BK_0+A_1e^{-s\Delta}+\int_{-\Delta}^{0}BK_1(\theta)e^{s\theta}d\theta-s\Big]^{-1}[E+BK_2]+[F+DK_2] \qquad (3.5)$$

In the sequel we shall use the following notation corresponding to (3.5):

$$C^c(s)\triangleq C_0+DK_0+C_1e^{-s\Delta}+\int_{-\Delta}^{0}DK_1(\tau)e^{s\tau}d\tau$$

$$A^c(s)\triangleq A_0+BK_0+A_1e^{-s\Delta}+\int_{-\Delta}^{0}BK_1(\tau)e^{s\tau}d\tau \qquad (3.6)$$

$$E^c(s)\triangleq E+BK_2$$

$$F^c(s)\triangleq F+DK_2$$

Also define:

$$A_q^c(\lambda)-\lambda\triangleq\begin{bmatrix}[A^c(\lambda)-\lambda] & \frac{d}{d\lambda}[A^c(\lambda)-\lambda] & \dots & \frac{1}{(q-1)!}\frac{d^{q-1}}{d\lambda^{q-1}}[A^c(\lambda)-\lambda]\\ 0 & [A^c(\lambda)-\lambda] & \dots & \vdots\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \dots & [A^c(\lambda)-\lambda]\end{bmatrix}$$

$$C_q^c(\lambda)\triangleq\begin{bmatrix}C^c(\lambda) & \frac{d}{d\lambda}C^c(\lambda) & \dots & \frac{1}{(q-1)!}\frac{d^{q-1}}{d\lambda^{q-1}}C^c(\lambda)\\ 0 & C^c(\lambda) & \dots & \vdots\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \dots & C^c(\lambda)\end{bmatrix}$$

$$E_q^c(\lambda) \triangleq \begin{bmatrix} E^c(\lambda) & \frac{d}{d\lambda}E^c(\lambda) & \cdots & \frac{1}{(q-1)!}\,\frac{d^{q-1}}{d\lambda^{q-1}}\,E^c(\lambda) \\ 0 & E^c(\lambda) & \cdots & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & E^c(\lambda) \end{bmatrix} \tag{3.7}$$

$$F_q^c(\lambda) \Delta \begin{bmatrix} F^c(\lambda) & \frac{d}{d\lambda}F^c(\lambda) & \cdots & \frac{1}{(q-1)!}\,\frac{d^{q-1}}{d\lambda^{q-1}}\,F^c(\lambda) \\ 0 & F^c(\lambda) & \cdots & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & F^c(\lambda) \end{bmatrix}$$

Similarly define $A_q(\lambda)$, $B_q(\lambda)$, $C_q(\lambda)$, $E_q(\lambda)$ and $F_q(\lambda)$ from (3.7), assuming the system is open loop (i.e. with $K_0=0$, $K_1=0$).

Also define:

$$\mathrm{Ker}(H-\lambda)_q \triangleq \mathrm{Ker} \underbrace{\begin{bmatrix} H-\lambda & -I & 0 \\ \vdots & & \vdots \\ 0 & 0 \;\cdots & -I \\ 0 & 0 \;\cdots & H-\lambda \end{bmatrix}}_{q \text{ blocks}} \tag{3.8}$$

$\forall\lambda \in \sigma(H)$ with q the multiplicity of λ in the minimal polynomial of H (1.2).

The following theorem gives a solution to the feedforward servomechanism problem with complete state measurement.

Theorem 1

The time delay feedforward servomechanism problem with complete state measurement is solvable for (1.1) with a control given by (3.2) iff the following conditions all hold:

i) $\mathrm{rank}\,[A_0+A_1e^{-\lambda\Delta}-\lambda \quad B] = n$

$\forall\lambda \in \sigma(A_0+A_1e^{-\lambda\Delta})$

ii)
$$\mathrm{rank}\begin{bmatrix} A_q(\lambda)-\lambda, & B_q(\lambda) \\ C_q(\lambda) \quad , & D_q(\lambda) \end{bmatrix} = \mathrm{rank}\begin{bmatrix} A_q(\lambda)-\lambda, & B_q(\lambda), & \begin{bmatrix} E_q(\lambda) \\ F_q(\lambda) \end{bmatrix} \mathrm{Ker}(H-\lambda)q \\ C_q(\lambda) \quad , & D_q(\lambda), & \end{bmatrix} \tag{3.9}$$

Remark 1

The solution for the time delay servomechanism problem with incomplete state measurement is not included for the sake of brevity; it may be found in [8].

4. COMPENSATORS FOR TIME DELAY SYSTEMS

In this section, a class of controllers for the time delay systems described by (1.1) is defined.

The following class of controllers is considered:

$$\begin{aligned}\dot{\eta} &= \Delta_0\eta + \Delta_1\eta(t-\Delta) + \int_{-\Delta}^{0}\Delta_2(\theta)\eta(t+\theta)d\theta + \int_{-\Delta}^{0}\Delta_3(\theta)\eta(t+\theta-\Delta)d\theta \\ &\quad + \Delta_4 y_m + \int_{-\Delta}^{0}\Delta_5(\theta)y_m(t+\theta)d\theta \qquad (4.1)\end{aligned}$$

Define the controller state space by $C^t \triangleq C([-\Delta,0];\ \mathbb{R}^t)$ and the state vector by $\eta_t(\theta) \triangleq \eta(t+\theta)$, $\theta\epsilon[-\Delta,0]$. In addition, from (4.1) by definition, $\Delta_0 \in \mathbb{R}^{t\times t}$, $\Delta_1 \in \mathbb{R}^{t\times t}$, $\Delta_4 \in \mathbb{R}^{t\times r_m}$ and

$$\begin{aligned}&\int_{-\Delta}^{0}\Delta_2(\theta)(\cdot)d\theta \ :\ C^t \to \mathbb{R}^t \\ &\int_{-\Delta}^{0}\Delta_3(\theta)(\cdot)d\theta \ :\ C^t \to \mathbb{R}^t\end{aligned} \qquad (4.2)$$

Define $y_{m_t}(\theta) \triangleq y_m(t+\theta)$, $\theta\epsilon[-\Delta,0]$; then the corresponding state space is C^{r_m} so that

$$\int_{-\Delta}^{0}\Delta_5(\theta)(\cdot)d\theta \ :\ C^{r_m} \to \mathbb{R}^t \qquad (4.3)$$

All the operators defined above are bounded operators.

The open loop augmented time delay system is therefore described by the following set of equations (augmenting (1.1) with (4.1) and assuming for simplicity that $D_m = 0$):

$$\begin{aligned}\dot{x}^a &= A_0^a x^a + A_1^a x^a(t-\Delta) + \int_{-\Delta}^{0}A_2^a(\theta)x^a(t+\theta)d\theta + \\ &\quad \int_{-\Delta}^{0}A_3^a(\theta)x^a(t-\Delta+\theta)d\theta + B^a u + E_0^a v + \int_{-\Delta}^{0}E_1^a(\theta)v(t+\theta)d\theta \\ e &= C_0^a x^a + C_1^a x^a(t-\theta) + D^a u + F^a v \qquad (4.4)\end{aligned}$$

where $\quad x^a \triangleq \begin{bmatrix} x \\ \eta \end{bmatrix}, \quad x^a \in \mathbb{R}^{\bar{n}\times\bar{n}}, \quad \bar{n} \triangleq n+t$

and $\quad x_t^a \triangleq x^a(t+\theta) \qquad \theta\in[-\Delta,0]$

$x_t^a \in C^{\bar{n}} \triangleq C([-\Delta,0];\ \mathbb{R}^{\bar{n}})$ and where

$$A_0^a \triangleq \begin{bmatrix} A_0 & 0 \\ \Delta_4 C_{0_m} & \Delta_0 \end{bmatrix}, \quad A_1^a \triangleq \begin{bmatrix} A_1 & 0 \\ \Delta_4 C_{1_m} & \Delta_1 \end{bmatrix}$$

$$A_2^a(\theta) \triangleq \begin{bmatrix} 0 & 0 \\ \Delta_5(\theta) C_{0_m} & \Delta_2(\theta) \end{bmatrix}$$

$$A_3^a(\theta) \triangleq \begin{bmatrix} 0 & 0 \\ \Delta_5(\theta) C_{1_m} & \Delta_3(\theta) \end{bmatrix}; \quad B_1^a \triangleq \begin{bmatrix} B \\ 0 \end{bmatrix}$$

$$E_0^a \triangleq \begin{bmatrix} E \\ \Delta_4 F_m \end{bmatrix}; \quad E_1^a(\theta) \triangleq \begin{bmatrix} 0 \\ \Delta_5(\theta) F_m \end{bmatrix} \tag{4.5}$$

$$C_0^a \triangleq (C_0 \quad 0); \quad C_1^a \triangleq (C_1 \quad 0)$$

$$D^a \triangleq D \qquad ; \quad F^a \triangleq F$$

The control input corresponding to (4.4) is now defined as follows:

$$u = K_0\eta + \int_{-\Delta}^{0} K_1(\theta)\eta(t+\theta)d\theta + K_2 y_m + \int_{-\Delta}^{0} K_3(\theta) y_m(t+\theta)d\theta \tag{4.6}$$

Substituting for y_m (1.1) one obtains

$$\begin{aligned} u = {} & K_0\eta + \int_{-\Delta}^{0} K_1(\theta)\eta(t+\theta)d\theta + K_2 C_{0_m} x + K_2 C_{1_m} x(t-\Delta) + \\ & \int_{-\Delta}^{0} K_3(\theta) C_{0_m} x(t+\theta)d\theta + \int_{-\Delta}^{0} K_3(\theta) C_{1_m} x(t-\Delta+\theta)d\theta + \\ & K_2 F_m v + \int_{-\Delta}^{0} K_3(\theta) F_m v(t+\theta)d\theta \end{aligned} \tag{4.7}$$

For the closed loop augmented system, the following is now defined:

$$A^a_{0c}(\theta) \triangleq \begin{bmatrix} A_0+BK_2C_{0_m} & BK_0 \\ \Delta_4C_{0_m} & \Delta_0 \end{bmatrix} \; ; \quad A^a_{1c} \triangleq \begin{bmatrix} A_1+BK_2C_{1_m} & 0 \\ \Delta_4C_{1_m} & \Delta_1 \end{bmatrix}$$

$$A^a_{2c}(\theta) \triangleq \begin{bmatrix} BK_3(\theta)C_{0_m} & BK_1(\theta) \\ \Delta_5(\theta)C_{0_m} & \Delta_2(\theta) \end{bmatrix} \; ; \quad A^a_{3c}(\theta) \triangleq \begin{bmatrix} BK_3(\theta)C_{1_m} & 0 \\ \Delta_5(\theta)C_{1_m} & \Delta_3(\theta) \end{bmatrix}$$

$$E^a_{0c} \triangleq \begin{bmatrix} E+BK_2F_m \\ \Delta_4F_m \end{bmatrix} \; ; \quad E^a_{1c}(\theta) \triangleq \begin{bmatrix} BK_3(\theta)F \\ \Delta_5(\theta)F_m \end{bmatrix} \qquad (4.8)$$

$$C^a_{0c} \triangleq (C_0+DK_2C_{0_m}, \; DK_0) \; ; \quad C^a_{1c} \triangleq (C_1+DK_2C_{1_m}, \; 0)$$

$$C^a_{2c}(\theta) \triangleq (DK_3(\theta)C_{0_m}, \; DK_1(\theta)) \; ; \quad C^a_{3c}(\theta) \triangleq (DK_3(\theta)C_{1_m}, \; 0)$$

$$F^a_{0c} \triangleq (F+DK_2F_m) \; ; \quad F^a_{1c}(\theta) \triangleq (DK_3(\theta)F_m) \qquad (4.9)$$

The closed loop augmented system is therefore now described as follows:

$$\dot{x}_a = A^a_{0c}x^a + A^a_{1c}x^a(t-\Delta) + \int_{-\Delta}^{0} A^a_{2c}(\theta)x^a(t+\theta)d\theta + \int_{-\Delta}^{0} A^a_{3c}(\theta)x^a(t-\Delta+\theta)d\theta + E^a_{0c}v + \int_{-\Delta}^{0} E^a_{1c}(\theta)v(t+\theta)d\theta$$

$$e = C^a_{0c}x^a + C^a_{1c}x(t-\Delta) + \int_{-\Delta}^{0} C^a_{2c}(\theta)x(t+\theta)d\theta + \int_{-\Delta}^{0} C^a_{3c}(\theta)x(t-\Delta+\theta)d\theta + F^a_{0c}v + \int_{-\Delta}^{0} F^a_{1c}(\theta)v(t+\theta)d\theta \qquad (4.10)$$

Applying now a Laplace transformation on (4.10), the transfer function $G^a_c(s)$ is obtained as follows:

$$e(s) = G^a_c(s)v(s) \qquad (4.11)$$

where

$$G^a_c(s) = [C^a_{0c} + C^a_{1c}e^{-s\Delta} + \int_{-\Delta}^{0} C^a_{2c}(\theta)e^{s\theta}d\theta + \int_{-\Delta}^{0} C^a_{3c}(\theta)e^{s(\theta-\Delta)}d\theta] \cdot [A^a_{0c} + A^a_{1c}e^{-s\Delta} + \int_{-\Delta}^{0} A^a_{2c}(\theta)e^{s\theta}d\theta + \int_{-\Delta}^{0} A^a_{3c}(\theta)e^{s(\theta-\Delta)}d\theta - s]^{-1} \cdot [E^a_{0c} + \int_{-\Delta}^{0} E^a_{1c}(\theta)e^{s\theta}d\theta] + [F^a_{0c} + \int_{-\Delta}^{0} F^a_{1c}(\theta)e^{s\theta}d\theta] \qquad (4.12)$$

It should be noted that the compensator (4.1) does not represent the most general compensator for time delay systems. In general, in a synthesis procedure, the degree of complexity of the compensator is limited by means of implementation; the structure chosen in (4.1) represents the simplest time delay controller with observer.

5. PERTURBATIONS ALLOWED IN SYSTEMS PARAMETERS

In this section, a class of perturbations which will be allowed in the development to follow, is defined. These perturbations represent variations of the system's parameters; the class of perturbations defined will be considered admissible if the stable closed loop time delay system remains stable for any perturbation in this class.

Definition 1

Given $M \in \mathbb{R}^{m\times n}$, define norm $\|M\| = \sum_{j=1}^{n} \sum_{i=1}^{m} |m_{ij}|$. For a given $\epsilon > 0$, the set Ω_ϵ is defined by

$$\Omega_\epsilon = \{M| \quad \|M\| \le \epsilon\} \tag{5.1}$$

Definition 2

Consider the system (2.5). Let $A_0 \to A_0 + \delta A_0$ (i.e. A_0 is replaced by $A_0 + \delta A_0$), $A_1 \to A_1 + \delta A_1$, $B \to B + \delta B$. If $(\delta A_0, \delta A_1, \delta B) \in \Omega_\epsilon$, then $(\delta A_0, \delta A_1, \delta B)$ represents a perturbation of the system's parameters.

In this development, the main interest is focused on maintaining stability of a perturbed system (2.3) with feedback

$$u = K_0 x(t) + \int_{-\Delta}^{0} K_1(\theta) x(t+\theta) d\theta \tag{5.2}$$

assuming that the closed loop system without perturbations is stable. For this purpose, we redefine the class of perturbations in $\mathbb{R}^n$ given in Definition 2 to be a class of perturbations in C^n. This is stated in the following Lemma.

Lemma 1

The class of perturbations in $\mathbb{R}^n$ given in Definitions 1, 2 induces in C^n a perturbation represented by a bounded operator.

The following theorem obtained from lemma 1 is the main result of this section. It gives a basic property for the class of admissible perturbations given in Definitions 1, 2.

Theorem 2

Given the system (2.3) with feedback (5.2) defined as a bounded operator $\mathcal{K}: C^n \to \mathcal{U}$ with bound M, assume the closed loop system is asymptotically stable. Then there exists an $\epsilon > 0$ so that for any $(\delta A_0, \delta A_1, \delta B) \in \Omega_\epsilon$ the perturbed closed loop system (i.e. $A_0 \to A_0 + \delta A_0$, $A_1 \to A_1 + \delta A_1$, $B \to B + \delta B$) is asymptotically stable.

Having defined the class of allowable perturbations in the system's parameters, the time delay servomechanism problem with parameter's perturbations is now defined and solved in the next section.

6. THE ROBUST TIME DELAY SERVOMECHANISM PROBLEM

The time delay servomechanism problem as defined in Section 3 is now considered with the additional requirement that output regulation occurs for all admissible perturbations satisfied by Theorem 2. This requirement creates the time delay servomechanism problem with perturbations. The problem is defined as follows: given the time invariant linear system described by (1.1), find necessary and sufficient conditions for the existence of a feedback denoted by $K: \mathcal{Y}_m \to \mathcal{U}$ so that the following conditions all hold.

a) the system is closed loop asymptotically stable

b) output regulation occurs, i.e.

$$\lim_{t\to\infty} e(t) = 0 \qquad \forall x_0 \in \mathcal{C}^n , \quad \forall v(0) \in \mathbb{R}^\ell$$

c) property b) holds for any admissible perturbation $(\delta A_0, \delta A_1, \delta B) \in \Omega_\epsilon$ which preserves closed loop stability (i.e. where $\epsilon > 0$ is chosen so that property a) holds).

In the above definition the feedback K may be a matrix of gains and/or a dynamical (controller).

In general for arbitrary admissible perturbations, conditions (3.9) do not hold. It therefore follows that a feedforward solution to the problem is not sufficient to provide a solution to the time delay servomechanism problem with perturbations.

It is now assumed that a controller (4.1) shall be applied to the system (1.1) defining the augmented time delay system. Denote the augmented state space $\mathcal{X}_a$ and the state trajectory vector $\binom{x}{\eta} \in \mathcal{X}_a$. Now the system is defined by (4.4)-(4.5) and the control input for the augmented time delay system is defined by (4.7). Following the introduction of the class of controllers (4.1) and the definition of the augmented time delay system, the following problem now arises. Given the time invariant linear system described by (1.1); find necessary and sufficient conditions for the existence of a controller (4.1) and a state feedback (4.6), so that there exists an $\epsilon > 0$ with the property that for all parameter perturbations $(\delta A_0, \delta A_1, \delta B) \in \Omega_\epsilon$, the following hold:

a) the augmented system is closed loop asymptotically stable

b) output regulation holds, i.e.

$$\lim_{t\to\infty} e(t) = 0 \qquad \forall x_t^a(0) \in \mathcal{C}^{n+t} \quad \text{and} \quad \forall v(0) \in \mathbb{R}^\ell$$

where x_t^a is the state vector of the infinitesimal representation [4] of the augmented time delay system.

This may be referred to as the robust time delay servomechanism problem. The controller (4.1) may be referred to as the robust time delay controller and is defined similarly to the robust controller as given in [3].

The definition of the equivalent external input w is now also introduced

$$w = C'v$$

$$\dot{v} = H v$$

where

$$H = \begin{pmatrix} 0 & 1 & \ldots\ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots\ldots & 1 \\ -p_1 & p_2 & \ldots\ldots & -p_p \end{pmatrix} ; \quad C' = (1,0,\ldots,0) \tag{6.1}$$

where H is the companion matrix corresponding to the minimal polynomial of H in (1.2)

The error system for the time delay system is now defined as follows:

$$\begin{aligned} \dot{x} &= A_0 x + A_1 x(t-\Delta) + Bu + I_n w \\ e &= C_0 x + C_1 x(t-\Delta) \\ y_m &= C_{m_0} x + C_{m_1} x(t-\Delta) \end{aligned} \tag{6.2}$$

where $w \in \mathbb{R}^n$ and every element of w is an equivalent external input.

In the next Theorem, it is shown that for a solution to the robust time delay servomechanism problem to exist, it is necessary for a feedforward controller to exist for the error system (6.2). For simplicity, it will be assumed in the development to follow that $D = D_m = 0$ (1.1).

Theorem 3

Consider (1.1), (1.2) and (4.1). A solution to the robust time delay servomechanism problem exists only if the feedforward time delay servomechanism problem is solvable for the error system (6.2).

The result obtained above allows us to further investigate the basic structural properties of a robust time delay controller. In essence, it is shown that the robust controller must consist of two parts:

a) a time delay servocompensator

b) a stabilizing compensator

A stabilizing compensator is defined to be an observer [10]; its sole purpose is to stabilize the augmented system (4.4). Now from (4.1), applying a Laplace Transformation, define

$$\begin{aligned} \Delta_0(s) &\triangleq \Delta_0 + \Delta_1 e^{-s\Delta} + \int_{-\Delta}^{0} \Delta_2(\tau) e^{s\tau} d\tau + \int_{-\Delta}^{0} \Delta_3(\tau) e^{s(\tau-\Delta)} d\tau \\ \Delta_1(s) &\triangleq \Delta_4 + \int_{-\Delta}^{0} \Delta_5(\tau) e^{s\tau} d\tau \end{aligned} \tag{6.3}$$

The following definitions will now be made:

$$\begin{aligned} A(s) &\triangleq A_0 + A_1 e^{-s\Delta} \\ C(s) &\triangleq C_0 + C_1 e^{-s\Delta} \\ C_m(s) &\triangleq C_{0_m} + C_{1_m} e^{-s\Delta} \end{aligned} \tag{6.4}$$

Definition 3

A time delay servo-compensator is defined as follows:

Let

$$r_1(\lambda) \triangleq \text{rank } C(\lambda)$$

$$r_2(\lambda) \triangleq \text{rank } [C(\lambda), \frac{d}{d\lambda} C(\lambda)]$$

$$\vdots$$

$$r_q(\lambda) \triangleq \text{rank } \left[C(\lambda), \ldots, \frac{1}{(q-1)!} \frac{d^{q-1}}{d\lambda^{q-1}} C(\lambda)\right]$$

Then $\Delta_0(s)$ (6.3) contains in its Jordan canonical structure $r_1(\lambda)$ blocks of size $q \times q$ for every $\lambda \in \sigma(H)$ where q is the multiplicity of λ in the minimal polynomial of H, $r_2(\lambda)-r_1(\lambda)$ blocks of size $(q-1)\times(q-1)$; ..... ; $r_q(\lambda)-r_{q-1}(\lambda)$ blocks of size 1×1.

In the development to follow, the following definition is also necessary.

Definition 4

The pseudo error $e(\lambda)$ for the system (1.1) is defined to be:

$$e(\lambda) \triangleq (C_0+C_1 e^{-\lambda\Delta})x + Fv$$

The pseudo measurement $y_m(\lambda)$ for the system (1.1) is defined to be:

$$y_m(\lambda) \triangleq (C_{0_m}+C_{1_m} e^{-\lambda\Delta})x + Fv$$

for all $\lambda \in \sigma(H)$.

Remark 2 If $C_1 = 0$ (1.1), then the pseudo errors are similar to the regulated errors.

In the following Theorem the structural characterization of the robust time delay controller is given.

Theorem 4

Given the system (1.1), a controller (4.1) is a robust time delay controller only if

a) $$\mathcal{R}[C_m(\lambda)]' \supset \mathcal{R}[C(\lambda)]' \qquad (6.5)$$

for all $\lambda \in \sigma(H)$, where $C_m(\lambda)$ and $C(\lambda)$ are given by (6.4).

b) it includes a time delay servocompensator according to the structure given by definition 3.

c) it includes a stabilizing compensator.

□

Remark 3 If $y_{ref} \neq 0$ the structure shown above is similar to that of a servocompensator [3] since $r_1(\lambda) = r$.

Remark 4 If $C_1 = 0$ the structure shown above is similar to that of a servocompensator since $r_1 = r_i(\lambda) = r$, $i = 1,2,\ldots,q$.

Corollary 1

(6.5) implies

$$\mathcal{R}[F_m(\lambda)]' \subset \mathcal{R}[F(\lambda)]' \tag{6.6}$$

It may be noted that the results in Corollary 1 and a) of Theorem 4 imply that the pseudo error must be the input to the time delay servocompensator.

In the following Theorem a complete solution to the robust time delay servomechanism problem is given.

Theorem 5

The robust time delay servomechanism problem is solvable iff the following conditions hold:

a) $$\operatorname{rank}\,[A_0+A_1e^{-\lambda\Delta}-\lambda,\quad B] = n \tag{6.7}$$

b) $$\operatorname{rank}\begin{bmatrix} A_0+A_1e^{-\lambda\Delta}-\lambda \\ \\ C_{0_m}+C_{1_m}e^{-\lambda\Delta} \end{bmatrix} = n \tag{6.8}$$

for all $\lambda \in \sigma(A_0+A_1e^{-\lambda\Delta})$

c) $$\operatorname{rank}\begin{bmatrix} A_q(\lambda)-\lambda & B_q(\lambda) \\ \\ C_q(\lambda) & 0 \end{bmatrix} = nq + \operatorname{rank}\,[C_q(\lambda)] \tag{6.9}$$

for all $\lambda \in \sigma(H)$ with q the multiplicity of λ in the minimal polynomial of H and $A_q(\lambda)$, $C_q(\lambda)$ and $D_q(\lambda)$ are defined in (3.7).

d) $$\mathcal{R}[C_m(\lambda)]' \supset \mathcal{R}[C(\lambda)]'$$

for all $\lambda \in \sigma(H)$.

7. CONCLUSIONS

In this paper the servomechanism problem for time delay systems (1.1) has been defined and studied. In Section 3 the time delay servomechanism problem is defined and solved (Theorem 1). In Section 4 a class of realizable compensators (4.1) is defined, and a class of admissible perturbations is defined in Section 5. The robust time delay servomechanism problem is then defined and solved in Section 6; necessary and sufficient conditions for a solution to exist to this problem are given in Theorem 5. In Theorems 3 and 4 structural properties of robust time delay controllers are given. In Theorem 3, it is shown that the robust problem is equivalent to a special type of feedforward time delay servomechanism problem. In Theorem 4 it is shown that the pseudo error must be an input to the robust time delay servocompensator.

8. REFERENCES

[1] Davison, E.J., "The Feedforward, Feedback and Robust Control of a General Servomechanism Problem, Part I", Control Systems Report No. 7305, Department of Electrical Engineering, U. of Toronto, 1973.

[2] Davison, E.J., "The Feedforward, Feedback and Robust Control of a General Servomechanism Problem, Part II", Control Systems Report No. 7306, Dept. of Electrical Engineering, U. of Toronto, 1973.

[3] Davison, E.J., Goldenberg, A., "Robust Control of a General Servomechanism Problem: the Servocompensator", Proc. I.F.A.C. 6th Triennial World Congress, 1975; Automatica, 1975, Vol. 11, No. 5, p. 461-4.

[4] Hale, J.K., Functional Differential Equations, New York Springler-Verlag, 1971.

[5] Dunford, N., Schwartz, J.T., Linear Operators - Part I, New York, Interscience Publishers, Inc., 1957.

[6] Krasovskii, N.N., "The Approximation of a Problem of Analytic Design of Controls in a System with Time-Lag", PRIKL. MATH. MEKH, 1964, Vol. 28, No. 4, p. 716-724.

[7] Vandevenne, H.F., "Controllability and Stabilizability Properties of Delay Systems", Proc. IEEE Conference on Decision and Control, 1972.

[8] Goldenberg, A., "On the Structure of the Linear Multivariable Servomechanism Problem Synthesis Problem", Ph.D. Thesis, September 1976, University of Toronto.

[9] Bhat, K.P.M., Koivo, H.N., "Modal Characterization of Controllability and Observability of Time Delay System", IEEE Trans. on Automatic Control, 1976, Vol. AC-21, No. 2, p. 292-3.

[10] Gressang, R.V., Lamont, G.B., "Observers for Systems Characterized by Semigroups", IEEE Trans. on Automatic Control, 1975, Vol. AC-20, No. 4, p. 523-528.

[11] Manitius, A., "Optimal Control of Hereditary Systems", Report CRM-462, Centre de Recherche Mathematiques, Université de Montréal, 1974.

[12] Bhat, K.P.M., Koivo, H.N., "An Observer Theory for Time Delay Systems", IEEE Trans. on Automatic Control, 1976, Vol. AC-21, No. 2, p. 266-269.

IDENTIFICATION OF MULTIVARIABLE SYSTEMS: A CRITICAL SURVEY

N.S. RAJBMAN
Institute of Control Sciences
USSR Academy of Sciences
Moscow, U.S.S.R.

N.K. SINHA
Department of Electrical Engineering
McMaster University
Hamilton, Canada

ABSTRACT

A critical survey is presented of the existing literature on the identification of multivariable systems from input-output data, along with some new directions for further research in the field. Emphasis is placed on realistic cases where the measurements are contaminated with noise.

INTRODUCTION

Although a considerable amount of work has been published on the identification of single-input single-output systems [1-5], relatively little has been done on the identification of multivariable systems. Pioneering work in this field was done by Gopinath [6] and Budin [7], but their methods are useful only when the measurement noise is negligible. Guidorzi [8] has presented an approach based on a canonical state-space representation which minimizes the number of parameters to be identified and also allows the determination of the structure of the system; but this method, again, is effective only for the case of noise-free observations. Gupta and Fairman [9] have proposed a method which does take the noise into account, but it is not possible to use their method for on-line identification. Sen and Sinha [10] and Sinha and Sen [11] have developed on-line methods for identification of multivariable systems which can take care of noise by introducing a noise model, but it is assumed that independent noise is added to each output without interaction. A more realistic approach is to assume interaction such that the noise at each output is fed back to the inputs of other variables, as proposed by Meerov, Pajziev and Rajbman [12]. They have suggested using partial correlation and dispersion functions for identification. An advantage of this approach is that it is applicable to nonlinear systems. However, in its present form, it is suitable only for off-line identification.

The object of this paper is to make a critical survey of the existing literature on the identification of linear multivariable systems, keeping in mind practical applications to real-life problems. New directions of current research in the field will also be discussed.

STATEMENT OF THE PROBLEM

Consider a linear time-invariant discrete-time system with r inputs and p outputs. The system can be represented by the state equations

$$\left.\begin{aligned} \underline{x}\,(k+1) &= A\,\underline{x}(k) + B\,\underline{u}(k) \\ \underline{y}\,(k) &= C\,\underline{x}(k) \end{aligned}\right\} \tag{1}$$

where $\underline{x}$ is the n-dimensional state vector, $\underline{y}$ is the p-dimensional output vector and $\underline{u}$ is the r-dimensional input vector. The system is thus specified by the matrix triple A, B and C.

The identification problem may then be stated as the determination of the matrices A, B and C from a given record of input-output observations. Since, in practice the observations are always contaminated with noise, a realistic method must take this fact into account. An alternative representation of the system defined by equations (1) is in terms of the transfer function matrix. Taking the z-transforms of the input and output vectors, we may write

$$\underline{Y}(z) = H(z)\ .\ \underline{U}(z) \tag{2}$$

where

$$\underline{Y}(z) = \begin{vmatrix} Y_1(z) \\ Y_2(z) \\ \vdots \\ Y_p(z) \end{vmatrix} = Z \begin{vmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_p(t) \end{vmatrix} \tag{3}$$

$$\underline{U}(z) = \begin{vmatrix} U_1(z) \\ U_2(z) \\ \vdots \\ U_r(z) \end{vmatrix} = Z \begin{vmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_r(t) \end{vmatrix} \tag{4}$$

and the symbol Z represents z-transformation for the given sampling period t_s, so that

$$z = \exp\ (s\ t_s) \tag{5}$$

In equation (2), H(z) is a rxp transfer function matrix, each element $h_{ij}(z)$ of which is a proper rational function of z.

It is well known that the transfer function matrix H(z) may be written as

$$H(z) = C\ (zI - A)^{-1}\ B \tag{6}$$

However, for a given H(z), one may obtain many sets of the matrices A, B and C. In all practical cases, it is sufficient to obtain a minimal dimensional realization of the system from the given set of input-output data, which is known to be both controllable and observable. The minimal realization of a transfer function matrix is not unique, however, unless a certain structure or a given canonical form is assumed.

Hence, the identification problem may be reformulated as the determination of the parameters of either the transfer function matrix H(z) or of the matrices A, B and C (assuming a given canonical form) from a record of the input-output data which may be contaminated with noise.

IDENTIFICATION FROM NOISE-FREE DATA

We shall first consider the case when the input-output records are free from noise, since this is the case which has received most attention in the literature. As mentioned in the introduction, the pioneering work in this field was done by Gopinath [6]. Budin [7] has presented a more direct procedure for realization in a well-defined structure, which uses a selector matrix similar to that of Gopinath. Although this method also gives the order of the system, besides determining the matrices A, B and C, it is computationally very involved.

Another approach, based on first determining the Markov parameters of the system from the input-output data, is computationally more straightforward. The procedure suggested in reference (11) will now be described briefly. Expanding the transfer function matrix in terms of the Markov parameters

$$H(z) = J_0\, z^{-1} + J_1\, z^{-2} + \ldots. \quad , \tag{7}$$

one may obtain the following matrix equation relating the input-output data,

$$[\underset{\sim}{y}(1) \quad \underset{\sim}{y}(2) \;\ldots] = [J_0 \quad J_1 \;\ldots] \left|\begin{matrix} \underset{\sim}{u}(0) & \underset{\sim}{u}(1) & \underset{\sim}{u}(2) & \ldots \\ \underset{\sim}{u}(-1) & \underset{\sim}{u}(0) & \underset{\sim}{u}(1) & \ldots \\ \underset{\sim}{u}(-2) & \underset{\sim}{u}(-1) & \underset{\sim}{u}(0) & \ldots \\ \vdots & \vdots & \vdots & \end{matrix}\right| \tag{8}$$

Since, for a stable system, one may assume that

$$\lim_{i\to\infty} J_i = 0, \tag{9}$$

it is possible to truncate the sequence of Markov parameters. One may then obtain a least-squares solution for the Markov parameters using the matrix pseudoinverse. From these estimates, one may then obtain the order of the system, as well as the minimum realization for the system in a canonical form using the procedure described by Rozsa and Sinha [13]. The main drawback with this approach is that a large number of Markov parameters must be estimated in order that an accurate representation of the system may be obtained.

An alternative approach, proposed by Sen and Sinha [10] uses the matrix pseudoinverse to estimate the parameters of the transfer function matrix. The state space realization may then be obtained using the method of reference [13]. The method can be used for on-line identification since the recursive algorithm requires the inversion of only a p x p matrix, where p is the number of outputs, which is usually small. The number of parameters to be estimated by this formulation is given by

$$q = n\ (pr+1) \tag{10}$$

which is higher than the number of unknown parameters in a canonical state space representation.

Guidorzi [8] has recently proposed a method which is very promising. It is based on a preliminary canonical structure identification, i.e., on the direct determination, from input-output sequences, of a set of invariant indexes completely describing the input-output structure of the system. Hence, besides these p indices, the number of system parameters to be estimated is only n(p+r), which is, generally, much smaller than the number q in equation (10). An important contribution of Guidorzi was to show that an input-output difference equation of the type

$$P(z)\ \underline{y}(k) = Q(z)\ \underline{u}(k) \tag{11}$$

can be obtained from the matrices A, B and C if these are in the row canonical form. In equation (11), P(z) and Q(z) are polynomial matrices in z, where z^{-1} is to be regarded as the unit delay operator. Furthermore, the polynomials in P(z) can be obtained directly from the A-matrix and vice versa. The polynomials in Q(z) depend on the elements of A as well as those of B, and knowing these, B can be uniquely determined. The details of the method can be obtained from reference [8]. This method, in the form presented by Guidorzi, however, is not suitable for on-line identification, nor is it very good in the presence of large amounts of measurement noise.

IDENTIFICATION FROM NOISY DATA

Although the various methods discussed in the previous section work very well when the input-output data contain negligible noise, they all cause problems with noisy data. The effect of noise can be particularly troublesome when one is determining the order or the structure of system.

If the order of the system is known, or if one is trying to fit a model of known order, it is possible to obtain unbiased least-squares estimates of the parameters. Some of the methods discussed in the literature will be briefly reviewed.

Guidorzi [8] has proposed obtaining least-squares estimates in a manner similar to the instrumental variable approach. Sen and Sinha [10] and Sinha and Sen [11] have proposed using the generalized least squares approach, where the system parameters and the noise model parameters are estimated in such a manner that the resulting residuals are uncorrelated and with zero mean. Whereas this method works well with single-input single-output systems it neglects the effect of interaction between the various outputs. A more realistic approach is to take such interaction into account by considering a noise model of the type shown in Figure 1 for a two-input two-output system. Thus for the two-input two-output linear plant of Fig. 1 whose overall characteristics are weighting functions of feed-forward and feedback loop in the presence of additive and noncorrelated noise we will have for the output variable $y_1(t)$

$$y_1(t) = \int_0^\infty g_{11}(\tau)u_1(t-\tau)d\tau - \int_0^\infty\int_0^\infty g_{11}(\tau)g_{12}(\tau')y_2(t-\tau-\tau')d\tau\, d\tau' + n_1(t) \quad (12)$$

A similar equation will hold for the output variable $y_2(t)$

$$y_2(t) = \int_0^\infty g_{22}(\tau)u_2(t-\tau)d\tau - \int_0^\infty\int_0^\infty g_{22}(\tau)g_{21}(\tau')y_1(t-\tau-\tau')d\tau\, d\tau' + n_2(t) \quad (13)$$

Let the additive noise $n_1(t)$ and $n_2(t)$ be uncorrelated among themselves and with the input signals $u_1(t)$ and $u_2(t)$. Then it is easy to show that optimal, in terms of the mean-root-square criterion, weighting functions $g_{11}(t)$ and $g_{12}(t)$, $g_{21}(t)$, $g_{22}(t)$ may be obtained by simultaneous solution of the following equations

$$K_{y_1u_1}(t) = \int_0^\infty g_{11}(\tau)K_{u_1u_1}(t-\tau)d\tau - \int_0^\infty\int_0^\infty g_{11}(\tau)g_{12}(\tau')K_{y_2u_1}(t-\tau-\tau')d\tau\, d\tau', \quad (14)$$

$$K_{y_1u_2}(t) = \int_0^\infty g_{22}(\tau)K_{u_2u_2}(t-\tau)d\tau - \int_0^\infty\int_0^\infty g_{22}(\tau)g_{21}(\tau')K_{y_1y_2}(t-\tau-\tau')d\tau\, d\tau', \quad (15)$$

$$K_{y_1y_2}(t) = \int_0^\infty g_{11}(\tau)K_{u_1u_2}(t-\tau)d\tau - \int_0^\infty\int_0^\infty g_{11}(\tau)g_{12}(\tau')K_{y_2y_2}(t-\tau-\tau')d\tau\, d\tau', \quad (16)$$

$$K_{y_2y_1}(t) = \int_0^\infty g_{22}(\tau)K_{u_2y_1}(t-\tau)d\tau - \int_0^\infty\int_0^\infty g_{22}(\tau)g_{21}(\tau')K_{y_2y_1}(t-\tau-\tau')d\tau\, d\tau', \quad (17)$$

where $K_{u_iu_i}(t)$, $K_{y_iy_i}(t)$ are correlation functions of the random input, $u_i(t)$ and output $y_i(t)$ functions respectively and $K_{y_iu_j}(t)$ are cross correlation functions of $y_i(t)$ and $u_j(t)$. Consequently, having found, from the knowledge of $u_1(t)$, $u_2(t)$, $y_1(t)$ and $y_2(t)$ the corresponding correlation and cross-correlation functions, one may solve eqns. (53-56) to determine optimal weighting functions $g_{11}(\tau)$ and $g_{22}(\tau)$ of feedforward loops and $g_{12}(\tau)$ and $g_{21}(\tau)$ of feedback loops of the two-variable system.

In solving the identification problem for a multi-dimensional system it is required to find "pure" or "actual" relations between random variables linked through feedforward and feedback channels. In this case a solution can be obtained by using partial characteristics. Let us first explain their use with a simple example. For the circuit of Fig. 1 the output variable y_1 is a function of the input u_1 and output y_2. Assume for symplicity that u_1, y_1 and y_2 are random. Then for a linear case the relation characteristics will be given as correlation coefficients $\tau_{y_1u_1}$, $\tau_{y_1y_2}$ and $\tau_{u_1y_2}$ obtained from the knowledge of experimental data obtained in normal functioning of the plant. However, the "pure", "actual" link of y_1 with u_1 along the feedforward channel and with y_2 along the feedback channel cannot be characterized by values of $\tau_{y_1u_1}$ or $\tau_{y_1y_2}$ because u_1, y_2 and y_1 are correlated; to put it differently, for a multidimensional system $\tau_{u_1y_2} \neq 0$. It is only partial coefficients of correlation that can give estimates of the relation between the quantitities in question. The partial coefficients are completely determined by the partial correlation coefficients $\tau_{y_1u_1}$, $\tau_{y_1y_2}$, $\tau_{u_1y_2}$. For our case the partial coefficients correlating y_1 and u_1 with the effect of y_2 disregarded, is equal to

$$\rho_{y_1u_1y_2} = \frac{\tau_{y_1u_1} - \tau_{y_1y_2}\tau_{u_1y_2}}{\sqrt{(1-\tau^2_{y_1y_2})(1-\tau^2_{u_1y_2})}} \tag{18}$$

Coefficients ρ for other cases may be written in a similar way.

Partial correlation functions permit simplification of the identification problem for a multivariable linear system. For instance, partial correlation functions lead to the following equation when the weighting functions $g_{11}(\tau)$ and $g_{22}(\tau)$ of a two-variable plant (Fig. 1) should be found

$$\rho_{y_1u_1}(t) = \int_0^\infty g_{11}(\tau)\rho_{u_1u_1}(t-\tau)\,d\tau, \tag{19}$$

$$\rho_{y_2u_2}(t) = \int_0^\infty g_{22}(\tau)\rho_{u_2u_2}(t-\tau)\,d\tau, \tag{20}$$

where $\rho_{u_1u_1}(t)$ and $\rho_{u_2u_2}(t)$ are the partial auto-correlation functions of random input functions $u_1(t)$ and $u_2(t)$ respectively; $\rho_{y_1u_1}(t)$ and $\rho_{y_2u_2}(t)$ are the partial cross-correlation functions of the random input-output functions $y_1(t)$, $u_1(t)$ and $y_2(t)$, $u_2(t)$ respectively.

It is easy to obtain the relation between the correlation and partial correlation functions

$$\rho_{u_2u_2}(\tau) = K_{u_2u_2}(\tau) - \int_0^\infty g_{21}(\tau)K_{y_1u_2}(\tau-\tau')d\tau', \tag{21}$$

$$\rho_{u_1u_1}(\tau) = K_{u_1u_1}(\tau) - \int_0^\infty g_{12}(\tau)K_{y_2u_1}(\tau-\tau')d\tau', \tag{22}$$

$$\rho_{y_1u_1}(\tau) = K_{y_1u_1}(\tau) - \int_0^\infty g_{21}(\tau)K_{y_1y_1}(\tau-\tau')d\tau', \tag{23}$$

$$\rho_{y_2u_2}(\tau) = K_{y_2u_2}(\tau) - \int_0^\infty g_{12}(\tau)K_{y_2y_2}(\tau-\tau')d\tau'. \tag{24}$$

Detail of a solution to the problem for two-variable case is presented in reference [15]. Extension to a multi-dimensional case should present no essential problems and it also is based on the use of partial random function correlation methods which yield linear models and partial dispersion functions which give nonlinear models.

DIRECTIONS OF FUTURE RESEARCH

From this survey, it appears that the problem of identification of multivariable systems from noisy input-output data is not completely solved yet. In particular, it is felt that on-line identification of multivariable systems requires considerable investigation.

In reference [16], Sinha and Kwong have developed a recursive method, based on the matrix pseudoinverse, which can be used to estimate the parameters (as well as the structural indices) following the method proposed by Guidorzi [8]. In the presence of noise, unbiased least-squares estimates of the parameters can be obtained by suitably modelling the noise. This is a subject for future research. Another possibility is the development of a recursive algorithm based on the correlation and dispersion methods proposed by Rajbman [15]. The frequency-domain approach proposed by Priestley et al. [14] can also be valuable for identification in the presence of noise.

ACKNOWLEDGEMENTS

This work was done while the second author visited the Institute of

Control Sciences under the exchange programme between the National Research Council of Canada and the USSR Academy of Sciences. The financial support from both of these agencies is gratefully acknowledged.

REFERENCES

[1] D. Graupe, "Identification of Systems", Van Nostrand Reinhold Company, 1972.

[2] A.P. Sage and J.L. Melsa, "System Identification", Academic Press, 1971.

[3] P. Eykhoff, "System Identification", John Wiley & Sons, 1974.

[4] G.N. Saridis, "Comparison of six on-line identification methods", Automatica, vol. 10, 1974, pp. 69-80.

[5] N.K. Sinha and A. Sen, "Critical evaluation of on-line identification methods", Proceedings of the Institution of Electrical Engineers, vol. 122, 1975, pp. 1153-1158.

[6] B. Gopinath, "On the identification of linear time-invarient systems from input-output data", Bell System Technical Journal, vol. 48, May-June 1969.

[7] M.A. Budin, "Minimal realization of discrete linear systems from input-output observations", IEEE Trans. on Automatic Control, vol. AC-16, 1971, pp. 395-401.

[8] R. Guidorzi, "Canonical structures in the identification of multivariable systems", Automatica, vol. 11, 1975, pp. 361-374.

[9] R.D. Gupta and F.W. Fairman, "An explicit approach to minimal realization from system Markov parameters", IEEE Trans. on Automatic Control, vol. AC-18, 1973, pp. 52-54.

[10] A. Sen and N.K. Sinha, "On-line estimation of the parameters of a multivariable system using matrix pseudoinverse", International Journal of Systems Science, vol. 7, 1976, pp. 461-471.

[11] N.K. Sinha and A Sen, "On-line identification of multivariable systems", Preprints IV IFAC Symp. on System Identification (Tbilisi, U.S.S.R.), Sept. 1976.

[12] M.V. Meerov, E. Pajziev and N.S. Rajbman, "Identification of stationary multivariable plants with the aid of partial correlation and variance functions", Automation and Remote Control, 1969, pp. 691-696.

[13] P. Rozsa and N.K. Sinha, "Efficient algorithm for irreducible realization of a rational matrix", International Journal of Control, vol. 20, 1974, pp. 739-751.

[14] M.B. Priestley, T. Subba Rao and H. Tong, "Applications of principal componentanalysis and factor analysis in the identification of multi-

variable systems", IEEE Trans. on Automatic Control, vol. AC-19, 1974, pp. 730-734.

[15] N.S. Rajbman, "Identification of multivariable systems by correlation and dispersion functions", to be presented at the IV IFAC Symp. on Multivariable Technological Systems (Fredericton, N.B.), July 1977.

[16] N.K. Sinha and Y.H. Kwong, "Recursive identification of the parameters of a multivariable system", ibid.

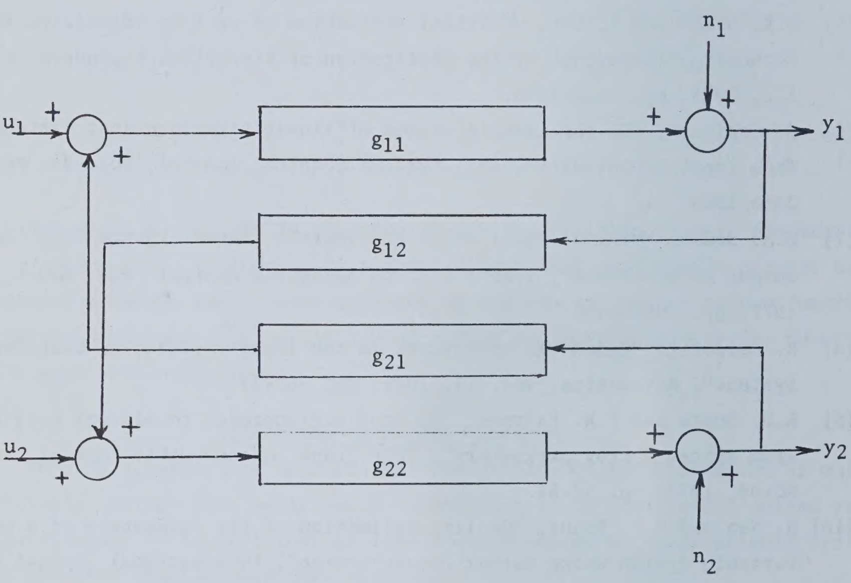


Figure 1: Block diagram of a two-input-two-output system with interactive noise.

IDENTIFICATION OF CERTAIN DYNAMIC SYSTEMS, FORCED BY DETERMINISTIC AND STOCHASTIC INPUTS, GIVEN THE INPUT AND NOISY OUTPUT MEASUREMENTS

EARL D. EYMAN AND EZEKIEL A. ADEDOTUN
Electrical Engineering Program, University of Iowa
Iowa City, Iowa 52242

ABSTRACT

The results presented in the paper of H. Weinert and E. Tse[1], "Structure determination and parameter identification for multivariable stochastic linear system," are expanded to certain systems with deterministic and additive stochastic inputs and noisy output measurement.

I. INTRODUCTION

Given the output or the input/output measurements, it is often required, either for control design or prediction, to obtain the best model for the dynamic system that has such input/output measurements. The determination of the structure coupled with the estimation of the parameter of the model forms the process of identification. Identification has been approached in many ways; of particular interest in this paper is the model considered by E. Tse and H. Weinert [1].

II. PROBLEM FORMULATION

E. Tse and H. Weinert [1] considered a stationary process, $\{\underline{y}(k)\}$, which is assumed to be the observation of a linear discrete dynamic system having the following dynamic and measurement equations:

$$\begin{aligned} \underline{x}(k+1) &= A\underline{x}(k) + \underline{e}(k) \\ \underline{y}(k) &= C\underline{x}(k) + \underline{\nu}(k) \end{aligned} \tag{1}$$

The set of equations (1) is replaced by the steady state innovation representation equations:

$$\begin{aligned} \hat{\underline{x}}(k+1|k) &= A\hat{\underline{x}}(k|k-1) + B\eta(k) \\ \underline{y}(k) &= C\hat{\underline{x}}(k|k-1) + \eta(k) \end{aligned} \tag{2}$$

where $\{\underline{\eta}(k)\}$ is the innovation sequence and B is steady state Kalman gain. Using the statistics of $\{\underline{y}(k)\}$, the structure and parameter of an invariant canonical representation of (2):

$$\{n, A_0, B_0, C_0, Q\} \tag{3}$$

is obtained.

In this paper we assume that $\{\underline{y}(k)\}$ is such that

$$\underline{y}(k) = \underline{y}_1(k) + y_2(k) \tag{4}$$

where

---

[1] Tse and Weiner [1]

$$\underline{x}_2(k + 1) = A\underline{x}_2(k) + G\underline{u}(k)$$
$$\underline{y}_2(k) = C\underline{x}_2(k) \tag{5}$$

$$\hat{\underline{x}}_1(k + 1|k) = A\hat{\underline{x}}_1(k|k - 1) + B\underline{\eta}(k)$$
$$\underline{y}_1(k) = C\hat{\underline{x}}(k|k - 1) + \underline{\eta}(k) \tag{6}$$

(A,G) is completely controllable

(A,B) is completely controllable

(A,C) is completely observable

A is stable

$\underline{u}(k)$ is bounded for all k

and $\underline{y}_1(k)$, $\underline{y}_2(k) \in R^m$, $\underline{x}(k) \in R^n$, $\underline{u}(k) \in R^r$. Given $\{\underline{y}(k)\}$, $\{\underline{u}(k)\}$, we determine the structure and estimate the parameter of an invariant canonical representation for the system.

Let

$$\underline{x}(k) = \hat{\underline{x}}_1(k|k - 1) + \underline{x}_2(k)$$

Then

$$\underline{x}(k + 1) = A[\hat{\underline{x}}_1(k|k - 1) + \underline{x}_2(k)] + G\underline{u}(k) + B\underline{\eta}(k)$$
$$= A\underline{x}(k) + G\underline{u}(k) + B\underline{\eta}(k) \tag{7}$$

$$\underline{y}(k) = \underline{y}_1(k) + y_2(k)$$
$$= C\underline{x}_1(k|k - 1) + \underline{\eta}(k) + C\underline{x}_2(k)$$
$$= C\underline{x}(k) + \underline{\eta}(k) \tag{8}$$

III. SEPARATION OF OUTPUT MEASUREMENT

Assume that $\underline{x}_2(0) = 0$, then from equation (5)

$$\underline{y}_2(k) = C \sum_{i=0}^{k-1} A^i G\underline{u}(k - i - 1) \tag{9}$$

For stable A and bounded $\underline{u}(k)$ for all k and for large k

$$\underline{y}_2(k) \approx C \sum_{i=0}^{\ell_1} A^i G\underline{u}(k - i - 1) \tag{10}$$

$$= \sum_{i=0}^{\ell_1} \theta_i \, \underline{u}(k - i - 1) \tag{11}$$

where $\theta_i \overset{\Delta}{=} CA^iG$. From (4) and (11) we have

$$\underline{y}(k) \approx \sum_{i=0}^{\ell} \theta_i \underline{u}(k - i - 1) + \underline{y}_1(k) \tag{12}$$

Assume that $\underline{y}_1(k)$ is stationary with an autoregressive model

$$\underline{y}_1(k) = - \sum_{i=1}^{\ell_2} \beta_i \underline{y}_1(k - i) + \underline{\alpha}(k) \tag{13}$$

then

$$\underline{y}(k) = \sum_{i=0}^{\ell_1} \theta_i \underline{u}(k - i - 1) + \underline{y}_1(k)$$

$$= \sum_{i=0}^{\ell_1} \theta_i \underline{u}(k - i - 1) - \sum_{i=1}^{\ell_2} \beta_i \underline{y}_1(k - i) + \underline{\alpha}(k) \tag{14}$$

Now, we need to obtain ℓ_1 and ℓ_2 and $\{\theta_i,\ i = 1,\ldots,\ell_1;\ \beta_j,\ j = 1,2,\ldots,\ell_2;\ R\}$ where R is the variance of the white noise process $\{\underline{\alpha}(k)\}$. Problems similar to the one defined above have been considered and solved:

Akaike's [2] considered the case where $\theta_i = 0$ for all i.

Kashyap [3] used conditional maximum likelihood.

Chan, et al. [4] solved the univariate problem.

In this paper we will use an algorithm which is a combination of those given by Van Den Boom, et al. [5] and Hsia [6].

Assuming N data points are available, we make the following definitions

$$y \overset{\Delta}{=} [\underline{y}^T(1), \underline{y}^T(2), \ldots, \underline{y}^T(N)]^T \tag{15}$$

$$y_1 \overset{\Delta}{=} [\underline{y}_1^T(1), \underline{y}_1^T(2), \ldots, \underline{y}_1^T(N)]^T \tag{16}$$

$$\psi_i \overset{\Delta}{=} [\underline{u}^T(i - 1), \ldots, \underline{u}^T(i - \ell_1)] \tag{17}$$

$$\psi^T \overset{\Delta}{=} [\psi_1^T, \psi_2^T, \ldots, \psi_N^T] \tag{18}$$

$$\theta \overset{\Delta}{=} [\theta_1^T, \theta_2^T, \ldots, \theta_{\ell_1}^T] \tag{19}$$

$$\beta \overset{\Delta}{=} [\beta_1^T, \beta_2^T, \ldots, \beta_{\ell_2}^T] \tag{20}$$

$$\phi_i \overset{\Delta}{=} [-\underline{y}_1^T(i-1), -\underline{y}_1^T(i-2), \ldots, -\underline{y}_1^T(i-\ell_2)] \tag{21}$$

$$\phi^T \triangleq [\phi_1^T, \phi_2^T, \ldots, \phi_N^T] \tag{22}$$

$$\alpha \triangleq [\underline{\alpha}^T(1), \underline{\alpha}^T(2), \ldots, \underline{\alpha}^T(n)]^T \tag{23}$$

From equations (15), (18) and (19)

$$y = \psi\theta + y_1 \tag{24}$$

From equations (16), (20), (22) and (23)

$$y_1 = \phi\beta + \alpha \tag{25}$$

From equations (24) and (25)

$$y = \psi\theta + \phi\beta + \alpha \tag{26}$$

$$= [\psi \quad \phi] \begin{bmatrix} \theta \\ \beta \end{bmatrix} + \alpha \tag{27}$$

(It should be remembered that ℓ_1 and ℓ_2 are still unknown.) To obtain ℓ_1 we use equation (24) and assume, for now, that $\{\underline{y}_1(k)\}$ is white noise.

$$y = \psi\theta + y_1 \tag{28}$$

Starting with $\ell_1 = 1$ we use the following steps to obtain ℓ_1^*

(a) Find $\hat{\theta}$ from (28)

$$\hat{\theta} = (\psi^T\psi)^{-1} \psi^T y \tag{29}$$

(b) Obtain $\hat{y}_1$ from equations (28) and (29)

$$\hat{y}_1 = y - \psi(\psi^T\psi)^{-1} \psi^T y \tag{30}$$

(c) Obtain det $1/N \; \hat{y}_1^T \; \hat{y}_1$

(d) Increase ℓ_1 by 1 and go back to a

(e) Let ℓ_1^* be ℓ_1 for the minimum in (c), to obtain ℓ_2 we use equation (25) together with steps (a) to (d) and $\hat{y}_1^*$ obtained in (e).

With ℓ_1^* and ℓ_2^* now available we focus our attention on getting an unbiased estimate of θ.

From equation (27)

$$\alpha = y - [\psi \quad \phi] \begin{bmatrix} \theta \\ \beta \end{bmatrix} \tag{31}$$

Let

$$J \triangleq \text{Trace} \frac{1}{N} \alpha^T \alpha \tag{32}$$

Our objective is to minimize J.

The least square estimates of $\hat{\theta}$ and $\hat{\beta}$ are obtained from

$$\begin{bmatrix} \hat{\theta} \\ \hat{\beta} \end{bmatrix} = \begin{pmatrix} \psi^T\psi & \psi^T\phi \\ \phi^T\psi & \phi^T\phi \end{pmatrix}^{-1} \begin{pmatrix} \psi^T y \\ \phi^T y \end{pmatrix} \tag{33}$$

using partitioned matrix

$$\hat{\theta} = (\psi^T\psi)^{-1}\,\psi^T y - (\psi^T\psi)^{-1}\,\psi^T\phi G^{-1}\,\phi^T\,Ny \tag{34}$$

$$\hat{\beta} = G^{-1}\,\phi^T\,Ny \tag{35}$$

where

$$N = I - \psi(\psi^T\psi)^{-1}\,\psi^T \tag{36}$$

$$G = \phi^T\,M\phi \tag{37}$$

From (34) and (35)

$$\hat{\theta} = (\psi^T\psi)^{-1}\,\psi^T y - (\psi^T\psi)^{-1}\,\psi^T\,\phi\hat{\beta} \tag{38}$$

$$\triangleq \hat{\theta}_{LS} - \hat{\theta}_{BIAS} \tag{39}$$

With equations (33) to (39) we can now use Hsia [6] algorithm.

Step 1: Assume $\hat{\beta} = 0$, compute

$$\hat{\theta}_{LS} = (\psi^T\psi)^{-1}\,\psi^T y \tag{40}$$

Step 2: Set $\hat{\theta} = \hat{\theta}_{LS}$ and generate $\hat{\underline{y}}_1(k) = \underline{y}(k) - \psi_k\hat{\theta}$. Obtain $\hat{\phi}$ from equations (21) and (22), compute $\hat{G} = \hat{\phi}^T\,M\phi$

Step 3: Compute

$$\hat{\beta} = \hat{G}^{-1}\,\hat{\phi}^T\,My \tag{41}$$

$$\hat{\theta}_{BIAS} = (\psi^T\psi)^{-1}\,\psi^T\,\hat{\phi}\hat{G} \tag{42}$$

update $\hat{\theta}$

$$\hat{\theta} = \hat{\theta}_{LS} - \hat{\theta}_{BIAS} \tag{43}$$

Step 4: Continue from Step 2 until the norm of the bias is very small. With $\hat{\theta}$ so obtained we can find $\{\hat{y}_1(k)\}$ by using

$$\hat{\underline{y}}_1(k) = \underline{y}(k) - \psi_k\hat{\theta} \tag{44}$$

We have now separated the output measurements into deterministic and stochastic parts.

IV. PARAMETER AND STRUCTURE OF THE INVARIANT CANONICAL FORM

$\{y_1(k)\}$ estimated in (44) is the stationary stochastic process. As in [1] we determine the structure and estimate the parameter of an invariant canonical form for the dynamic and measurement equations (6). We refrain from duplicating the derivation; however, for clarification, we repeat some equations from [1]. We assume the following order or rows of the observability matrix:

$$\begin{aligned} &C_1', \; C_1'A, \; \dots \\ &C_2', \; C_2'A, \; \dots \\ &C_m', \; C_m'A, \; \dots \end{aligned} \tag{45}$$

where C_i' is the ith row of C. Let p_i, $i = 1,\dots,m$ be the structural indices. Then by definition of $\{p_i\}_{i=1}^{m}$ there exists $\{\beta_{ijk}\}$, such that

$$C_i' A^{p_i} = \sum_{j=1}^{i} \sum_{k=0}^{p_j-1} \beta_{ijk} C_j' A^k \;, \qquad \text{if} \qquad p_i > 0 \;. \tag{46}$$

$$C_i' = \sum_{j=1}^{i-1} \sum_{k=0}^{p_j-1} \beta_{ijk} C_j' A^k \;, \qquad \text{if} \qquad p_i = 0 \tag{47}$$

for $i = 1,2,\dots,m$. A_0 and C_0 are obtained from $\{p_i,\beta_{ijk}\}$:

$$A_0 = \begin{bmatrix} A_{11} & & \\ \vdots & 0 & \\ A_{m1} & \dots & A_{mm} \end{bmatrix} \tag{48}$$

$$\underset{(p_i \times p_i)}{A_{ii}} = \begin{bmatrix} I & & 0 \\ \beta_{ii0} & & \beta_{ii1},\dots,\beta_{iip_i-1} \end{bmatrix} \tag{49}$$

$$\underset{\substack{(p_i \times p_j) \\ i>j}}{A_{ij}} = \begin{bmatrix} & 0 & \\ \beta_{ij0} & \dots & \beta_{ijp_j-1} \end{bmatrix} \tag{50}$$

$$C_{0_i}' = [0,\dots,1,0,\dots,0] \;, \qquad p_i > 0 \tag{51}$$

where the 1 in C_{0_i}' is in column $1 + p_1 + p_2 + \dots + p_i - 1$ and

$$C'_{0_i} = [\beta_{i10},\ldots,\beta_{i1p_1-1},\ldots,\beta_{ii-1p_{i-1}} - 1,0,\ldots,0] \qquad (52)$$

$$\text{if } p_i = 0$$

The parameters in equations (49), (50), (51) and (52) are estimated from the statistics of $\{\underline{y}_1(k)\}$ [1].

As in [1] we estimate B and Q by using the iteration procedure

$$\Sigma_0 = 0$$

$$\begin{aligned} Q_i &= R(0) - C\,\Sigma_i\,C' , && i = 0,1,\ldots \\ B_i &= (S - A\,\Sigma_i\,C')Q_i^{-1} , && i = 0,1,\ldots \\ \Sigma_{i+1} &= A\,\Sigma_i\,A' + B_i\,Q_i\,B_i' , && i = 0,1,\ldots \end{aligned} \qquad (53)$$

where S can be computed from

$$\begin{bmatrix} r_{1j}(1) \\ \vdots \\ r_{1j}(p_1) \\ \vdots \\ r_{mj}(p_m) \end{bmatrix} = \begin{bmatrix} C_1' \\ \vdots \\ C_1' A^{p_1-1} \\ \vdots \\ C_m' A^{p_m-1} \end{bmatrix} S_j , \qquad j = 1,\ldots,m \qquad (54)$$

$$R(0) = \frac{1}{N} \sum_{k=1}^{N} \underline{y}_1(k)\,\underline{y}'(k) \qquad (55)$$

$r_{ij}(\sigma)$ is the i,jth element of $R(\sigma)$ and S_j is the jth column of S. Σ is the covariance matrix of the state.

Now we obtain G_0 from equation (19)

$$\hat{\theta} = \left[\hat{\theta}_1^T, \hat{\theta}_2^T, \ldots, \hat{\theta}_{\ell_1^*}^T\right] \qquad (56)$$

and

$$\hat{\theta}_i = C_0\,A_0^i\,G_0 \qquad (57)$$

NUMERICAL EXAMPLE

We consider the following single-unit/two-output system.

$$\underline{x}(k+1) = \begin{bmatrix} 0.0 & 1.0 \\ -.1 & -.7 \end{bmatrix} \underline{x}(k) + \begin{bmatrix} 1.0 \\ 0 \end{bmatrix} \underline{u}(k) + \begin{bmatrix} 1.0 \\ .125 \end{bmatrix} e(k)$$

$$\underline{y}(k) = \begin{bmatrix} 1.0 & 0 \\ 0.0 & 1.0 \end{bmatrix} \underline{x}(k) + \begin{bmatrix} 1.0 & 0 \\ 0 & 1.0 \end{bmatrix} \underline{\nu}(k)$$

$$E[e(k)\ e(\ell)] = 2.0\ \delta_{k\ell}$$

$$E[\underline{\nu}(k)\ \underline{\nu}(\ell)] = I\ \delta_{k\ell}$$

(I is identity matrix)

$$E[\underline{x}(0)\ x^t(0)] = \begin{bmatrix} .2 & 0 \\ 0 & .1 \end{bmatrix}$$

$$u(k) = 1.0 + \left(1. + \frac{11k}{100}\right) \exp -\frac{k}{1000}$$

With 500 data points of input and outputs we obtain the following results.

$$\underline{x}(k+1) = \begin{bmatrix} 0. & 1.0 \\ \beta_1 & \beta_1 \end{bmatrix} \underline{x}(k) + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} u(k) + \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \underline{\eta}(k)$$

$$\underline{y}(k) = \begin{bmatrix} 1.0 & 0 \\ \gamma_1 & \gamma_2 \end{bmatrix} \underline{x}(k) + \underline{\eta}(k)$$

$$E[\underline{\eta}(k)\ \underline{\eta}(\ell)] = R\ \delta_{k\ell}$$

	Experimental Value	Actual Value
β_1	- .187	- .1
β_2	-1.19	- .7
b_1	1.2	1.0
b_2	0.2	0.
γ_1	0.	0.
γ_2	1.	1.
a_1	1.0	.167
a_3	1.0	+ .074
a_2	0.0	- .009
a_4	.4	.05

CONCLUSION

We have shown how to determine the structure and estimate the parameter of an invariant canonical form for certain dynamic systems, forced by deterministic and stochastic inputs, given the input and noisy output measurements.

REFERENCES

1. E. Tse and H. Weinert, "Structure determination and parameter identification for multivariable stochastic linear system," IEEE Trans. Auto. Cont. (1975).

2. H. Akaike, "Autoregressive model fitting for control," Ann. Inst. Statist. Math. (1971).

3. R. L. Kashyap, "Maximum likelihood identification of stochastic linear system," IEEE Trans. Auto Cont. (1970).

4. C. Chan, J. Harris and E. Wellstead, "An order-testing criterion for mixed autoregressive moving average processes," Int. Jour. Cont. (1974).

5. A. Van Den Boom and A. Van Den Enden, "The determination of the order of process and noise dynamics," IFAC Identification and System Parameter Estimation (1973).

6. T. C. Hsia, "On least squares algorithms for system parameter identification," IEEE Trans. Aut. Cont. (1976).

OPTIMAL INPUT SIGNALS FOR IDENTIFICATION OF DISTRIBUTED PARAMETER SYSTEMS: A FREQUENCY-DOMAIN APPROACH

B. KUSZTA
Institute of Control and Electronics
Technical University of Warsaw
Warsaw, Poland

N.K. SINHA
Department of Electrical Engineeri
McMaster University
Hamilton, Canada

ABSTRACT

A frequency-domain approach is presented for the design of optimal input signals for identifying the parameters of a linear distributed parameter system. It is assumed that the set of possible values of the parameters of the unknown system is given.

INTRODUCTION

In an earlier paper [1], it was shown that the optimal input signal for the identification of a linear lumped-parameter system has its power spectral density proportional to $|\Delta G(j\omega)|^2$ where $\Delta G(j\omega)$ is defined as the maximum deviation in the transfer function $G(j\omega)$ as the parameter vector varies between its limiting values. This result was derived for the requirement that the input signal should satisfy the constraint that its power spectral density function be square integrable. On the other hand, if the input signal was to satisfy an energy constraint, then the optimal input was shown to be a sinusoid at frequency corresponding to the peak of $|\Delta G(j\omega)|^2$. Both of these results were based on the work of Gagliardi [2], who has shown that "for a finite set of parameters the input vectors whose output set has the largest perimeter when plotted in vector space maximize the probability of identifying the correct parameters".

The object of this paper is to extend the earlier results to the case of systems with distributed parameters. As an example, the one-dimensional heat-conduction system is considered. The proposed method is intuitively appealing, and very meaningful for engineers. Furthermore, it can be implemented into practice by designing a suitable staircase function as in a recent paper [3].

STATEMENT OF THE PROBLEM

Let the given system (Fig. 1) be described on R = (0,L) x (0,T) by the partial differential equation

$$\frac{\partial u(t,x)}{\partial t} = L_x \{a_i, u(t,x)\} \tag{1}$$

with initial condition and boundary conditions at x=0 and X=L, given by

$$\left.\begin{aligned} u_o &= u(x,\ t_o) \quad \text{for } t = t_o \\ u\Big|_{x=0} &= f(t) \\ u\Big|_{x=L} &= 0 \end{aligned}\right\} \tag{2}$$

In equation (1) L_x represents a linear space operator, a_i $i=1,2,...p$ represents a discrete set of possible values for the parameter a, which will be assumed to be equally a priori. The problem is to select the optimum input f(t) from an admissible set F for determining, from the observation $u(t,x_o)$ (in chosen point x_o) of the output, which member of the set a_i is the true value of a. It will be assumed that the observations of the output are contaminated with zero - mean white Gaussion noise.

A FREQUENCY-DOMAIN SOLUTION TO THE PROBLEM

As mentioned in the introduction the optimal input signal is the signal whose output has the largest perimeter when plotted in vector space.

For the case where i = 2 (it can be easily extended to the more general case when i = p) this requirement can be expressed in the following form

$$I = \int_0^\infty \left| u_1(t,x_o) - u_2(t,x_o) \right|^2 dt = \max \tag{3}$$

where $u_{1/2}(t,x_o)$ relates to $u(t,x_o)$ in (1) for $a_i = a_1$ and for $a_i = a_2$ respectively.

The solution of the system (1) with initial and boundary condition is obtained following Tichonow and Samarski [4],

$$\begin{aligned} u(a_i,t,x_o) &= \int_0^L G(a_i,t,t_o,x_o,\xi)\ u(\xi,t_o)d\xi + \int_0^t G'(a_i,\tau,t_o,x_o)\ f(\tau)d\tau \\ &= u'(a_i,t,x_o) + u''(a_i,t,x_o) \end{aligned} \tag{4}$$

where

G(.) is Green's function of the system (1),

G'(.) is "boundary" Green's function,

u'(.) output signal component connected with non-zero initial conditions, and

u''(.) output signal component connected with external boundary input signal.

Since during the identification only the u''(.) component can be affected by the input signal, the input selection criterion should be rewritten in the following form

$$I = \int_0^\infty \left| u_1''(t,x_o) - u_2''(t,x_o) \right|^2 dt = \max \tag{5}$$

Combining (4), (5) and taking into account Parseval's formula, we have

$$\begin{aligned} I &= \int_0^\infty \left| G'(a_1,t_o,x_o,j\omega) - G'(a_2,t_o,x_o,j\omega) \right|^2 \phi(\omega)\, d\omega \\ &= \int_0^\infty \left| \Delta H(t_o,x_o,j\omega) \right|^2 \phi(\omega)\, d\omega \end{aligned} \tag{6}$$

where

$$H(.) = \int_0^\infty G'(a_i,t_o,x_o,\tau)\, e^{j\omega\tau}\, d\tau \tag{7}$$

and $\phi(\omega)$ is the power spectral density of a bounded input signal. It is easily shown [1] that I is maximized if

$$\phi(\omega) = K \left| \Delta H(t_o,j\omega) \right|^2 \tag{8}$$

where K is a constant.

EXAMPLE

Consider the following one-dimensional heat-conduction system

$$\frac{\partial u(x,t)}{\partial t} = a^2 \frac{\partial^2 u(x,t)}{\partial x^2} \tag{9}$$

with boundary and initial conditions

$$\begin{aligned} u(x,0) &= 0 \\ u(0,t) &= f(t) \\ u(\ell,t) &= 0 \end{aligned} \tag{9a}$$

The Green's function for the system (9) has a form [2]

$$G(a,t,\tau,x,\xi) = \frac{2}{\ell} \sum_{n=1}^{\infty} \sin\left(\frac{n\pi x}{\ell}\right) \sin\left(\frac{n\pi}{\ell}\xi\right) \exp\left[- \frac{n^2\pi^2}{\ell^2} a^2(t-\tau)\right] \tag{10}$$

Since the eqn. (10) is time-invariant the substitution of

$$\lambda = t - \tau$$

gives the Green's function of the form

$$G(a,\lambda,x,\xi) = \frac{2}{\ell} \sum_{n=1}^{\infty} \sin\left(\frac{n\pi}{\ell}x\right) \sin\left(\frac{n\pi}{\ell}\xi\right) \exp\left(- \frac{n^2\pi^2}{\ell^2} a^2\lambda\right) \tag{11}$$

The "boundary" Green's function takes the form [4]

$$G'(a,\lambda,x) = a^2 \left. \frac{\partial G(a,\lambda,x,\xi)}{\partial \xi} \right|_{\xi=0} \tag{12}$$

The explicit form of boundary Green's function has a form

$$G'(.) = \frac{2\pi}{\ell^2} a^2 \sum_{n=1}^{\infty} n\, e^{-\frac{n^2\pi^2a^2}{\ell^2}\lambda} \sin\left(\frac{n\pi}{\ell}x\right) \tag{13}$$

Finally

$$u''(a,x,t) = u(a,x,t) = \frac{2\pi a^2}{\ell^2}\int_0^t \sum_{u=1}^{\infty} n\, e^{-\frac{n^2\pi^2a^2}{\ell^2}(t-\tau)} \sin\frac{n\pi}{\ell}x \; f(\tau)\, d\tau \tag{14}$$

Eqn. (14) can be obtained directly by applying Laplace transformation to (9) and (9a). Transforming both sides of (9) and (9a) one can get

$$a^2\,\frac{d^2u(x,p)}{dx^2} = p\, u(x,p) \tag{15}$$

and

$$u(x,p) = C_1(p)\, e^{\frac{x\sqrt{p}}{a}} + C_2(p)\, e^{-\frac{x\sqrt{p}}{a}}$$

$$\begin{aligned} u(0,p) &= f(p) \\ u(\ell,p) &= 0 \end{aligned} \tag{16}$$

From (16)

$$C_1(p) = -f(p)\,\frac{e^{-\frac{\ell\sqrt{p}}{a}}}{\mathrm{sh}\frac{\ell\sqrt{p}}{a}} \tag{17}$$

$$C_2(p) = f(p)\,\frac{e^{\frac{\ell\sqrt{p}}{a}}}{\mathrm{sh}\,\frac{\ell\sqrt{p}}{a}} \tag{18}$$

and finally

$$u(x,p) = f(p)\,\frac{\mathrm{sh}\frac{(\ell-x)\sqrt{p}}{a}}{\mathrm{sh}\,\frac{\ell\sqrt{p}}{a}} \tag{19}$$

In the time domain [5]

$$u(x,t) = \frac{2\pi a^2}{\ell^2}\left\{f(t) * \sum_{n=1}^{\infty} n \sin\left(\frac{n\pi x}{\ell}\right) e^{-\frac{n^2\pi^2a^2}{\ell^2}t}\right\} \tag{20}$$

where * denotes convolution.

Taking into account only one term (n=1) of boundary Green's function the power spectral density of input signal takes the form

$$\phi(\omega) = |\Delta H(j\omega,t_o)|^2 = \left| \int^{\infty} a^2 \left.\frac{\Delta\partial G(a,t,x_o,\partial\xi)}{\partial\xi}\right|_{\xi=0} e^{-j\omega t}\, dt \right|^2$$

$$= \left| \frac{2\pi^2}{\ell^2} \sin\left(\frac{\pi x_o}{\ell}\right) \left| \frac{1}{\frac{\pi^2 a_1^2}{\ell^2} + j\omega} - \frac{1}{\frac{\pi^2 a_2^2}{\ell^2} + j\omega} \right| \right|^2 = \left| \frac{2\pi^2}{\ell^2} \sin\left(\frac{\pi x_o}{\ell}\right) \frac{\frac{\pi^2}{\ell^2}\left(a_2^2 - a_1^2\right)}{\left(\frac{\pi^2 a_1^2}{\ell^2} + j\omega\right)\left(\frac{\pi^2 a_2^2}{\ell^2} + j\omega\right)} \right|$$

For $\ell=\pi$, $x_o=\frac{\ell}{2}$, $a_1=1.1$, $a_2=0.9$

$$\phi(\omega) = k \frac{\omega^2}{(0.98 - \omega^2)^2 + 2.02\omega^2} \qquad (21)$$

The shape of power density (21) is shown in fig. 2a. One of the possible input signals can be selected as

$$f(t) = F^{-1}[\psi(j\omega)] \qquad (22)$$

where

$$\psi(j\omega)\ \psi(-j\omega) = \phi(\omega) \qquad (23)$$

In the given example

$$f(t) = 3.025\ e^{-1.21t} - 2.025\ e^{-0.81t} \qquad (24)$$

Function f(t) can be approximated by Walsh function series [6]. Eight term approximation to f(t) has the form

$$f(t) = -.179\psi_o + .613\psi_1 + .298\psi_{10} - .004\psi_{11} + .149\psi_{100} + - .001\psi_{101} - .001\psi_{110} + .011\psi_{111} \qquad (25)$$

The power spectral density of the approximated function by (25) is shown in fig. 2(b).

CONCLUSIONS

A simple procedure for obtaining the optimal input signal is obtained, which maximizes the probability of obtaining the correct estimates of the parameters. The frequency-domain criterion thus obtained is intuitively appealing and easy to apply.

An example of a one-dimensional heat-conduction probelm shows how the method can be used in a practical case.

REFERENCES

1) B. Kuszta, N.K. Sinha, "Optimal input signals for system identification: A frequency-domain approach", Proc. 18th Midwest Symp. on Circuits and Systems, (Montreal, Canada), August 1975, pp. 149-152.

2) B. Kuszta and N.K. Sinha, "Optimal staircase signal for system identification", Proc. 13th Annual Allerton Conf. on Circuits and Systems, Monticello, Illinois), October 1975, pp. 829-838.

3) R.M. Gagliardi, "Input selection for parameter identification in discrete systems", IEEE Trans. vol. AC-12, 1967, pp. 597-599.

4) A.H. Tichonow, A.A. Samarski, "Equations of mathematical physics" (in Russian), G.I.T.T.L., Moscow, 1951.

5) J. Hladik, "La transformation de Laplace'a plusieurs variables", Massou et Cie Editeurs, Paris, 1969.

6) M.S. Corrington, "Solution of differential and integral equations with Walsh functions", IEEE Trans. vol. CT-20, No. 5, September 1973, pp. 470-476.

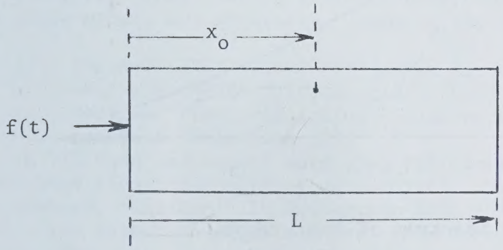


Figure 1: A Distributed-Parameter System

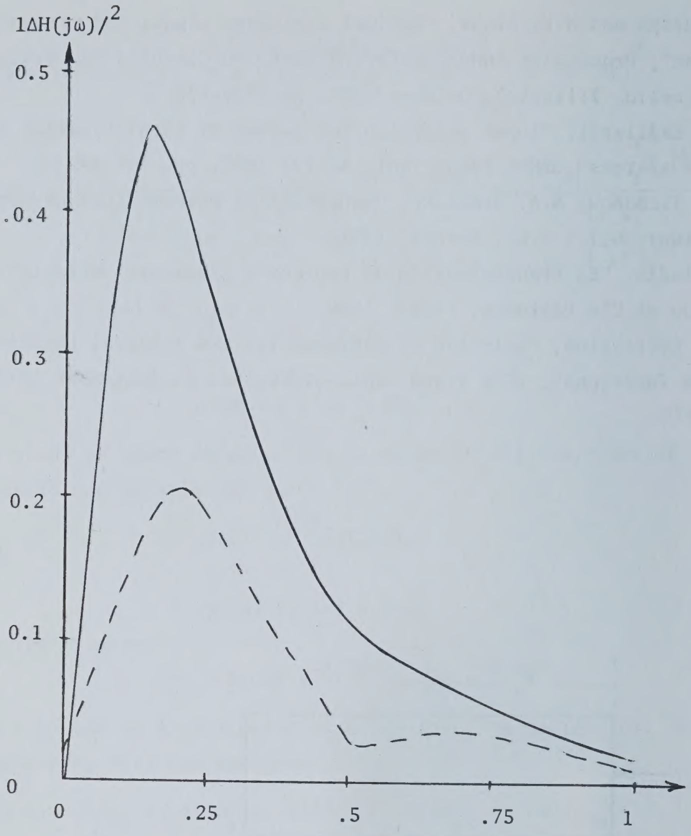


Figure 2: Power Spectrum of Input Signal

(a) desired
(b) Walsh function approximation

AN EXTENSION OF MODIFIED FINAL PREDICTION ERROR CRITERION TO TESTING THE ORDERS OF MULTIVARIATE ARMA PROCESSES AND MULTIVARIABLE SYSTEMS

EZEKIEL A. ADEDOTUN and EARL D. EYMAN
Electrical Engineering Program
University of Iowa
Iowa City, Iowa 52242

ABSTRACT

The Modified Final Prediction Error Criterion [5] is an order-testing procedure for univariate autoregressive-moving-average model for a process and the model for single input/output system. In this paper we extend this procedure to testing the orders of multivariate autoregressive-moving-average model for certain processes and the model of some multiple-input/multiple-output.

INTRODUCTION

In [4] Valis showed how certain given state variable models can be reduced to models involving only the input/output measurements and a stochastic process. Also, it is well known that a given stochastic process can be modeled by Autoregressive-moving-Average model (ARMA). A complete identification of either of the models consists of:

a. Determination of the orders of the model
b. Estimation of the parameters of the model.

Different order testing criteria have been proposed; some are subjective while others are objective. Some of the objective testing criteria are:

a. Akaike's (FPE) Final Prediction Error Scheme [1], [2];
b. Akaike's (MFPE) Multiple Final Prediction Error Scheme [3]; and
c. Modified Final Prediction Scheme proposed by C. Chan, J. Harris and P. E. Wellstead [5].

In [5] Chan and others note that FPE becomes subjective when the order is far less than the number of data points. The given algorithmic equations [5], however, consider only univariate ARMA model and single input/output model. The objective of this paper, therefore, is to extend [5] to a multivariate ARMA model and a multiple input/output model for a dynamic system.

The paper is divided into two sections: In Section I, the algorithmic equation for a multivariate ARMA model is derived. In Section II, the algorithmic equation in I is extended to a multiple input/output model.

SECTION I

We consider a multivariable model:

$$\underline{y}(k) = - \sum_{i=1}^{N_y} A_i \, \underline{y}(k-i) + \sum_{i=o}^{N_u} B_i \underline{u}(k-i)$$

$$+ \sum_{i=1}^{N_c} C_i \, \underline{z}(k-i) + \underline{z}(k) \qquad (1)$$

where $\{\underline{y}(k)\}$, $\{\underline{u}(k)\}$ and $\{\underline{z}(k)\}$ are the output, the input and the noise sequences.

N_y, N_u, N_C are unknown orders.

$\{\underline{z}(k)\}$ is gaussian white noise

$$E[\underline{z}(k)] = 0$$

$$E[\underline{z}(k)\ \underline{z}^T(\ell)] = R\delta_{k\ell} \tag{2}$$

The initial condition is zero. Let θ be matrix of the coefficients A_i's, B_i's, C_i's.

Let

$$\bar{\underline{y}}(k) \triangleq E\{\underline{y}(k)|\underline{y}(i),\ \underline{u}(i),\ 1 \le i \le k-1,\ \theta\} \tag{3}$$

$$\hat{\bar{\underline{y}}}(k) \triangleq E\{\underline{y}(k)|y(i),\ \underline{u}(i),\ 1 \le i \le k-1,\ \hat{\theta}\}$$

where $\hat{\bar{\underline{y}}}(k)$ is MLE of $\bar{\underline{y}}(k)$

$$p\{\underline{y}(k)|y(i),\ u(i),\ 1 \le i \le k-1,\ \theta\} \sim N(\bar{y}(k),\ R) \tag{4}$$

Assume N data points are available.

Let $y^k \triangleq \{\underline{y}(k),\ k = 1,\ \ldots,\ N\}$.

$$\ell np\{y^k|y(i),\ u(i),\ 1 \le i \le N-1,\ \theta\} \triangleq$$

$$L^N(\theta,R) = \ln(\frac{1}{2\pi})^{\frac{mN}{2}} + \ln|R|^{-\frac{N}{2}} + \ln\exp -\frac{1}{2}\sum_{k=1}^{N}(\underline{y}(k) - \bar{\underline{y}}(k))^T R^{-1}$$

$$(\underline{y}(k) - \bar{\underline{y}}(k)) \tag{5}$$

$$= \text{const} - \frac{N}{2}\ell n|R| - \frac{1}{2}\sum_{k=1}^{N}(\underline{y}(k) - \bar{\underline{y}}(k))^T R^{-1}$$

$$(\underline{y}(k) - \bar{\underline{y}}(k)) \tag{6}$$

Let $\hat{\theta}$ be the MLE of θ and the cost function be $v(\hat{\theta})$, where

$$v(\hat{\theta}) = \frac{N}{2}\det\frac{1}{N}\sum_{k=1}^{N}(y(k) - \hat{\bar{\underline{y}}}(k))(\underline{y}(k) - \hat{\bar{\underline{y}}}(k))^T \tag{7}$$

$$= \frac{N}{2}|\hat{R}| \tag{8}$$

where $\hat{\bar{\underline{y}}}(k)$ and $\hat{R}$ are MLE of $\bar{\underline{y}}(k)$ and R and are:

$$\hat{\bar{\underline{y}}}(k) = -\sum_{i=1}^{N_y}\hat{A}_i\ \underline{y}(k-i) + \sum_{i=1}^{N_u}\hat{B}_i\underline{u}(k-i) + \sum_{i=1}^{N_z}\hat{C}\hat{\underline{z}}(k-i) \tag{9}$$

$$\hat{R} = \frac{1}{N}\sum_{k=1}^{N}(\underline{y}(k) - \hat{\bar{\underline{y}}}(k))\ (\underline{y}(k) - \hat{\bar{\underline{y}}}(k)^T \tag{10}$$

It is well known that the minimum of (7) with respect to $\hat{\theta}$ is equivalent to the maximum of $L^N(\hat{\theta},R)$ (See [8]).

Since MLE is asymptotically consistent, then

$$\lim_{N\to\infty} \hat{\theta}^N = \theta_o \quad \text{and} \quad \lim_{N\to\infty} \hat{R}_N = R$$

with probability one.

Also $\sqrt{N}\,(\theta_o - \hat{\theta}^N)$ is asymptotically normally distributed with mean zero and variance $-NL^{-1}_{\theta\theta}(\theta_o)$. For large N, $L^{-1}_{\theta\theta}(\theta_o)$ can be replaced by $L^{-1}_{\theta\theta}(\hat{\theta}^N)$.

From eqs. (5) and (7)

$$\frac{\partial L^N(\hat{\theta},R)}{\partial\theta} = \sum_{\gamma,\delta} \frac{\partial L^N(\hat{\theta},R)}{\partial\left\langle \underline{y}(k) - \hat{\bar{\underline{y}}}(k)\right\rangle_{\gamma,\delta}} \cdot \frac{\partial\left\langle \underline{y}(k) - \hat{\bar{\underline{y}}}(k)\right\rangle_{\gamma,\delta}}{\partial\,\theta} \tag{11}$$

$$= \sum_{\gamma,\delta} - \sum_{k=1}^{N} R^{-1}\left\langle \underline{y}(k) - \hat{\bar{\underline{y}}}(k)\right\rangle_{\gamma,\delta} \frac{\partial\left\langle \underline{y}(k) - \hat{\bar{\underline{y}}}(k)\right\rangle_{\gamma,\delta}}{\partial\theta}$$

$$\frac{\partial v(\hat{\theta})}{\partial\theta} = \sum_{\gamma,\delta} \frac{\partial v(\hat{\theta})}{\partial\left\langle \underline{y}(k) - \hat{\bar{\underline{y}}}(k)\right\rangle_{\gamma,\delta}} \cdot \frac{\partial\left\langle \underline{y}(k) - \hat{\bar{\underline{y}}}(k)\right\rangle_{\gamma,\delta}}{\partial\,\theta}$$

$$= \frac{N}{2}|\hat{R}|\cdot\hat{R}^{-1} \sum_{\gamma,\delta}^{N} \frac{2}{N}\left\langle \underline{y}(k) - \hat{\bar{\underline{y}}}(k)\right\rangle_{\gamma,\delta} \frac{\partial\left\langle \underline{y}(k) - \hat{\bar{\underline{y}}}(k)\right\rangle_{\gamma,\delta}}{\partial\theta}$$

($|\hat{R}|\ \hat{R}^{-1}$ = I NOISE IS WHITE)

$$\frac{\partial v(\hat{\theta})}{\partial\theta} = \sum_{\gamma,\delta}\sum_{k=1}^{N}\left\langle \underline{y}(k) - \hat{\bar{\underline{y}}}(k)\right\rangle_{\gamma,\delta} \frac{\partial\left\langle \underline{y}(k) - \hat{\bar{\underline{y}}}(k)\right\rangle_{\gamma,\delta}}{\partial\theta} \tag{12}$$

From (11) and (12)

$$L^N_\theta\ (\hat{\theta}^N) = -\ R^{-1}\ v\theta(\hat{\theta}^N) \tag{13}$$

then

$$E[(\theta_o - \hat{\theta}^N)(\theta_o - \hat{\theta}^N)^T] = -\ L^{-1}_{\theta\theta}(\hat{\theta}^N) = Rv^{-1}_{\theta\theta}(\hat{\theta}^N) \tag{14}$$

ARMA MODEL

Now consider the model

$$\underline{y}(k) = -\sum_{i=1}^{N_y} A_i\underline{y}(k - i) + \sum_{i=1}^{N_z} B_i\underline{z}(k - i) + \underline{z}(k) \tag{15}$$

$$\hat{\bar{\underline{y}}}(k) = -\sum_{i=1}^{N_y} \hat{A}_i\underline{y}(k - i) + \sum_{i=1}^{N_z} \hat{B}_i\underline{z}(k - 1) + \underline{z}(u)$$

$$\hat{R} = \frac{1}{N} \sum_{k=1}^{N} (\underline{y}(k) - \hat{\bar{\underline{y}}}(k))(\underline{y}(k) - \hat{\bar{\underline{y}}}(k))^T$$

$$= \frac{1}{N} \sum_{k=1}^{N} \hat{\underline{z}}(k)\ \hat{\underline{z}}^T(k) \tag{16}$$

where

MLE of $\underline{z}(k)$ is $\hat{\underline{z}}(k)$

$$v(\hat{\theta}) = \frac{N}{2} \det \frac{1}{N} \sum_{k=1}^{N} \hat{\underline{z}}(k)\ \hat{\underline{z}}^T(k) \tag{17}$$

Minimizing $v(\hat{\theta})$ w. r. t. $\hat{\theta}$

$$\frac{\partial v(\hat{\theta})}{\partial \hat{\theta}} = 0 \tag{18}$$

$$\frac{\partial v(\hat{\theta})}{\partial \hat{A}_\ell} = \sum_{\gamma,\delta} \frac{\partial v(\hat{\theta})}{\partial \langle \underline{y}(k) - \hat{\bar{\underline{y}}}(k) \rangle_{\gamma,\delta}} \cdot \frac{\partial \langle \underline{y}(k) - \hat{\bar{\underline{y}}}(k) \rangle_{\gamma,\delta}}{\partial \hat{A}_\ell} \tag{19}$$

$$\frac{\partial (\underline{y}(k) - \hat{\bar{\underline{y}}}(k))}{\partial \langle \hat{A}_\ell \rangle} = J\ \underline{y}(k - \ell) - \hat{B}\ \frac{\partial \hat{\underline{z}}(k - \ell)}{\partial \hat{A}_\ell} \tag{20}$$

$$\frac{\partial \langle \underline{y}(k) - \hat{\bar{\underline{y}}}(k) \rangle}{\partial \hat{A}_\ell} = K\underline{y}^T(k - \ell) - \hat{B}\ \frac{\partial \langle \hat{\underline{z}}(k - \ell) \rangle}{\partial \hat{A}_\ell} \tag{21}$$

then

$$\frac{\partial v(\hat{\theta})}{\partial \hat{A}_\ell} = \sum_{\gamma,\delta} \sum_{k=1}^{N} \langle \underline{y}(k) - \hat{\bar{\underline{y}}}(k) \rangle_{\gamma,\delta} \Big[k_{\gamma,\delta} \underline{y}^T(k - \ell)$$

$$- \hat{B}\ \frac{\partial \langle \hat{\underline{z}}(k - \ell) \rangle_{\gamma,\delta}}{\partial \hat{A}_\ell} \Big]$$

$$= \sum_{k=1}^{N} (\underline{y}(k) - \hat{\bar{\underline{y}}}(k))\underline{y}^T(k - \ell)$$

$$- \sum_{k=1}^{N} \hat{B} \langle \hat{\underline{z}}(k) \rangle_{\gamma,\delta}\ \frac{\partial \langle \hat{z}(k - \ell) \rangle_{\gamma,\delta}}{\partial \hat{A}_\ell}$$

$$= \sum_{k=1}^{N} [\underline{y}(k) - \hat{\bar{\underline{y}}}(k)]\underline{y}^T(k - \ell) = 0 \tag{22}$$

Substitute for $\hat{\underline{z}}(k)$

(where $\hat{\underline{z}}(k) = - \sum_{i=1}^{N_y} (A_i - \hat{A}_i)\underline{y}(k - i) + \sum_{i=1}^{N_z} B_i z(k - i) + z(k) - \sum_{i=1}^{N_z} B_i \underline{z}(k-i))$

and rearrange

$$\begin{aligned}\sum_{k=1}^{N} \underline{z}(k)y^T(k - \ell) = \sum_{k=1}^{N} \Bigg[& \sum_{i=1}^{N_y} (A_i - \hat{A}_i)\underline{y}(k - i)\underline{y}^T(k - \ell) \\ & + \sum_{i=1}^{N_z} \hat{B}_i\hat{\underline{z}}(k - i)\underline{y}^T(k - \ell) \\ & - \sum_{i=1}^{N_z} B_i\underline{z}(k - i)y^T(k - \ell)\Bigg]\end{aligned} \tag{23}$$

also

$$\sum_{k=1}^{N} \hat{\underline{z}}(k)\hat{\underline{z}}^T(k - \ell) = 0 \quad \text{for} \quad 1 \neq 0 \tag{24}$$

substitute for $\underline{z}(k)$ and rearrange

$$\begin{aligned}\sum_{k=1}^{N} \underline{z}(k)\hat{\underline{z}}^T(k - \ell) = \sum_{k=1}^{N} \Bigg[& \sum_{i=1}^{N_y} (A_i - \hat{A}_i)\underline{y}(k - i)\hat{\underline{z}}^T(k - \ell) \\ & + \sum_{i=1}^{N_z} \hat{B}_i\hat{\underline{z}}(k - i)\hat{z}^T(k - \ell) \\ & - \sum_{i=1}^{N_z} B_i\underline{z}(k - i)\hat{\underline{z}}(k - \ell)\Bigg]\end{aligned} \tag{25}$$

$$\hat{R} = \frac{1}{N} \sum_{k=1}^{N} \hat{\underline{z}}(k)\hat{\underline{z}}^T(k) \tag{26}$$

substitue for $\hat{z}(k)$

$$\begin{aligned}\hat{R} = \frac{1}{N} \sum_{k=1}^{N} \Bigg[& -\sum_{i=1}^{N_y} (A_i - \hat{A}_i)\underline{y}(k - i)\hat{\underline{z}}^T(k) \\ & + \sum_{i=1}^{N_z} \{B_i\underline{z}(k - i)\hat{\underline{z}}^T(k) - \hat{B}_i\hat{\underline{z}}(k - i)\hat{\underline{z}}^T(k)\} \\ & + \underline{z}(k)\hat{z}^T(k)\Bigg]\end{aligned} \tag{27}$$

Noting that

$$\sum_{i=1}^{N_z} B_i\hat{\underline{z}}(k - i)\hat{z}^T(k) = 0 \tag{28}$$

and $$\sum_{k=1}^{N} \hat{\underline{z}}(k)\underline{y}^T(k - \ell) = 0 \tag{29}$$

Then eq. (27) becomes

$$\hat{R} = \frac{1}{N}\sum_{k=1}^{N}\left[\underline{z}(k)\hat{z}^T(k) + \sum_{j=1}^{N_z} B_j\underline{z}(k - j)\hat{\underline{z}}^T(k)\right] \tag{30}$$

Substitute for $\hat{\underline{z}}^T(k)$ in (30)

$$\begin{aligned}\hat{R} = {} & \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_z} B_j\underline{z}(k - j)\hat{\underline{z}}^T(k) \\ & - \frac{1}{N}\sum_{k=1}^{N}\underline{z}(k)\sum_{j=1}^{N_y}\underline{y}^T(k - j)(A_j - \hat{A}_j)^T + \frac{1}{N}\sum_{k=1}^{N}\underline{z}(k)\underline{z}^T(k) \\ & - \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_z}\underline{z}(k)\hat{\underline{z}}^T(k - j)\hat{B}^T\end{aligned} \tag{31}$$

$$\begin{aligned}\hat{R} = R & + \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_z} B_j\underline{z}(k - j)\hat{\underline{z}}^T(k) - \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_z}\underline{z}(k)\hat{z}^T(k - j)\hat{B}^T \\ & - \frac{1}{N}\sum_{k=1}^{N}\underline{z}(k)\sum_{j=1}^{N_y} y^T(k - j)(A_j - A_j)^T\end{aligned} \tag{32}$$

Substitute for $\sum_{k=1}^{N}\underline{z}(k)\underline{y}^T(k - j)$

$$\begin{aligned}\hat{R} = R & + \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_z} B_j\underline{z}(k - j)\hat{\underline{z}}^T(k) - \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_z}\underline{z}(k)\hat{\underline{z}}^T(k - j) \\ & (\hat{B}_j{}^T) - \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_y}\sum_{i=1}^{N_y}(A_i - \hat{A}_i)\underline{y}(k - i)\underline{y}^T(k - j)(A_j - \hat{A}_j)^T \\ & - \sum_{k=1}^{N}\sum_{j=1}^{N_y}\sum_{i=1}^{N_z}\hat{B}_i\underline{z}(k - i)y^T(k - j)(A_j - \hat{A}_j)^T \\ & + \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_y}\sum_{i=1}^{N_y} B_i\underline{z}(k - i)\underline{y}^T(k - j)(A_j - \hat{A}_j)^T\end{aligned} \tag{33}$$

Since $\hat{R}$ is a consistent estimate of R, all terms on the right are small compared with R. The following approximations are made.

$$[B_i\,\underline{z}(k - i) - \hat{B}_i\hat{\underline{z}}(k - i)] \approx (B_i - \hat{B}_i)\hat{\underline{z}}(k - i) \tag{34}$$

$$\sum_{k=1}^{N} \hat{\underline{z}}(k)z^T(k-1) \approx \sum_{k=1}^{N} \underline{z}(k)\hat{z}^T(k-i) \tag{35}$$

$$\begin{aligned}\hat{R} \approx R &+ \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_y}\sum_{i=1}^{N_z}(B_i-\hat{B}_i)\hat{\underline{z}}(k-i)\underline{y}^T(k-j)(A_j-\hat{A}_j)^T \\ &- \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_y}\sum_{i=1}^{N_y}(A_i-\hat{A}_i)\underline{y}(k-i)\underline{y}^T(k-j)(A_j-\hat{A}_j)^T \\ &+ \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_z}B_j\underline{z}(k-j)\hat{\underline{z}}^T(k) \\ &- \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_z}\underline{z}(k)\hat{\underline{z}}^T(k-j)\,\hat{B}_j^T\end{aligned} \tag{36}$$

Substitute (25) into (36)

$$\begin{aligned}\hat{R} \approx R &+ \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_y}\sum_{i=1}^{N_z}(B_i-\hat{B}_i)\hat{\underline{z}}(k-i)\underline{y}^T(k-j)(A_j-\hat{A}_j)^T \\ &- \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_y}\sum_{i=1}^{N_y}(A_i-\hat{A}_i)\underline{y}(k-i)\underline{y}^T(k-j)(A_j-\hat{A}_j)^T \\ &+ \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_z}B_j\underline{z}(k-j)\hat{\underline{z}}^T(k) \\ &- \sum_{k=1}^{N}\sum_{j=1}^{N_z}\sum_{i=1}^{N_y}(A_i-\hat{A}_i)\underline{y}(k-i)\hat{z}^T(k-j)\hat{B}_j^T \\ &+ \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_z}\sum_{i=1}^{N_z}\hat{B}_i\hat{\underline{z}}(k-i)\hat{z}^T(k-j)\hat{B}_j^T \\ &+ \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N_z}\sum_{i=1}^{N_z}B_i\underline{z}(k-i)\hat{z}^T(k-j)\hat{B}_j^T\end{aligned} \tag{37}$$

Using the approx. $\hat{\underline{z}}(k) \simeq \underline{z}(k)$ and substitute for $\hat{\underline{z}}^T(k)$ and using (18) and taking the expectation of both sides.

$$\begin{aligned}\hat{R} &\approx R + 2RB_1 - RA_1 - RC_1 \\ R &\approx \hat{R}[I + 2B_1 - A_1 - C_1]^{-1}\end{aligned} \tag{38}$$

where

$$A_1 = \sum_{i=1}^{N_y}\sum_{j=1}^{N_y}\left(\frac{\partial v}{\partial A_i\,\partial A_j}\right)^{-1} R_{\underline{yy}}(i-j)$$

$$B_1 = \sum_{i=1}^{N_z} \sum_{j=1}^{N_y} \left(\frac{\partial^2 v}{\partial A_i \partial B_j} \right)^{-1} R_{\underline{y}\hat{\underline{z}}} (i - j)$$

$$C_1 = \sum_{i=1}^{N_z} \sum_{j=1}^{N_z} \left(\frac{\partial^2 v}{\partial B_i \partial B_j} \right)^{-1} R_{\hat{z}\hat{z}} (i - j)$$

$$= \sum_{i=1}^{N_z} \left(\frac{\partial^2 v}{\partial B_i \partial B_i} \right)^{-1} R_{\hat{z}\hat{z}} (0)$$

and

$$R_{AB}(i - j) = \frac{1}{N-q} \sum_{k=q+1}^{N} A_{k-i} B^T_{k-j}$$

$$q = \max (i, j)$$

Now

$$\hat{\underline{z}}(k) = \sum_{i=1}^{N_y} (A_i - \hat{A}_i)\underline{y}(k - i) + \sum_{i=1}^{N_z} [B_i \underline{z}(k - i) - \hat{B}_i \hat{\underline{z}}(k - i)] + \underline{z}(k)$$

$$E[\hat{\underline{z}}(k)\hat{\underline{z}}^T(k)] = E\left[- \sum_{i=1}^{N_y} (A_i - \hat{A}_i)\underline{y}(k - i) + \sum_{i=1}^{N_z} \{\hat{B}_i \hat{\underline{z}}(k - i) - \hat{B}_i \hat{\underline{z}}(k - i)\} + \underline{z}(k)\right]\left[- \sum_{j=1}^{N_y} \underline{y}^T(k - j) (A_j - \hat{A}_j)^T + \sum_{j=1}^{N_z} \underline{z}^T(k - j)B_j^T - \hat{\underline{z}}^T(k - j)\hat{B}_j + \underline{z}^T(k)\right] \tag{39}$$

with approximations (35) and (34)

$$E[\hat{\underline{z}}(k)\hat{\underline{z}}^T(k)]$$

$$= E\left[- \sum_{i=1}^{N_y} (A_i - \hat{A}_i)\underline{y}(k - 1) + \sum_{i=1}^{N_z} (\hat{B}_i - \hat{B}_i)\hat{\underline{z}}(k - i) + \underline{z}(k)\right]$$

$$\left[- \sum_{j=1}^{N_y} \underline{y}^T(k - j)(A_j - \hat{A}_j)^T + \sum_{j=1}^{N_z} \hat{\underline{z}}^T(k - j)(B_j - \hat{B}_j)^T + \underline{z}^T(k)\right]$$

$$= E \sum_{i=1}^{N_y} \sum_{j=1}^{N_y} (A_i - \hat{A}_i)\underline{y}(k - i)\underline{y}^T(k - j)(A_j - \hat{A}_j)^T$$

$$- 2E\left[\sum_{i=1}^{N_z} \sum_{j=1}^{N_y} (B_i - \hat{B}_i)\hat{\underline{z}}(k - i)\underline{y}^T(k - j)(A_j - \hat{A}_j)^T \right]$$

$$+ E\left[\sum_{i=1}^{N_z}\sum_{j=1}^{N_z}(B_i - \hat{B}_i)\hat{\underline{z}}(k-i)\hat{\underline{z}}^T(k-j)(B_j - \hat{B}_j)^T\right] + R$$

$$+ E\left[\sum_{i=1}^{N_z}(B_i - \hat{B}_i)\hat{\underline{z}}(k-i)z^T(k)\right]$$

$$\approx R + E\left[\sum_{i=1}^{N_y}\sum_{j=1}^{N_y}(A_i - \hat{A}_i)\underline{y}(k-i)\underline{y}^T(k-j)(A_j - \hat{A}_j)^T\right]$$

$$- 2E\left[\sum_{i=1}^{N_z}\sum_{j=1}^{N_y}(B_i - \hat{B}_i)\hat{\underline{z}}(k-i)\underline{y}(k-j)(A_j - \hat{A}_j)^T\right]$$

$$+ E\left[\sum_{i=1}^{N_z}\sum_{j=1}^{N_z}(B_i - \hat{B}_i)\hat{\underline{z}}(k-i)\hat{\underline{z}}^T(k-i)(B_j - \hat{B}_j)^T\right] \qquad (40)$$

using eq. (18)

$$E[\hat{\underline{z}}(k)\hat{\underline{z}}^T(k)] = R + RA_1 - 2RB_1 + RC_1$$

$$= R[I + A_1 - 2B_1 + C_1]$$

$$= R[I + 2B_1 - A_1 - C_1]^{-1}[I + A_1 - 2B_1 + C_1] \qquad (41)$$

For the best model the norm of $E[\hat{\underline{z}}(k)\hat{\underline{z}}^T(k)]$ is small.

SECTION II

Consider the multiple input/output model:

$$\underline{y}(k) = -\sum_{i=1}^{N_y} A_i\underline{y}(k-i) + \sum_{i=0}^{N} B_i\underline{u}(k-i) + \sum_{i=1}^{N_z} C_i\,\underline{z}(k-i) + \underline{z}(k) \qquad (42)$$

$$\hat{\hat{\underline{y}}}(k) = -\sum_{i=1}^{N_y}\hat{A}_i\underline{y}(k-i) + \sum_{i=0}^{N_u}\hat{B}_i\underline{u}(k-i)$$

$$+ \sum_{i=1}^{N_z}\hat{C}_i\hat{\underline{z}}(k-i) \qquad (43)$$

Then the estimated variance of the process is

$$\hat{R} = \frac{1}{N}\sum_{k=1}^{N}(\underline{y}(k) - \hat{\hat{\underline{y}}}(k))(\underline{y}(k) - \hat{\hat{\underline{y}}}(k))^T$$

and the cost function

$$v(\hat{\theta}) = \frac{N}{2}\det\frac{1}{N}\sum_{k=1}^{N}(\underline{y}(k) - \hat{\hat{\underline{y}}}(k))(y(k) - \hat{\hat{\underline{y}}}(k))^T \qquad (44)$$

By differentiating (44) and following the procedure in Section I it can be shown that

$$R \approx R[I - A_2 - B_2 - C_2 + 2D_2 + 2E_2]^{-1}$$

where

$$A_2 = \sum_{i=1}^{N_y} \sum_{j=1}^{N_y} \left(\frac{\partial^2 v}{\partial A_i \partial A_j}\right)^{-1} R_{\underline{yy}} (i - j)$$

$$B_2 = \sum_{i=0}^{N_u} \sum_{j=0}^{N_u} \left(\frac{\partial^2 v}{\partial B_i \partial B_j}\right)^{-1} R_{\underline{uu}} (i - j)$$

$$C_2 = \sum_{i=1}^{N_z} \left(\frac{\partial^2 v}{\partial C_i \partial C_i}\right)^{-1} R_{\underline{\hat{z}}\underline{\hat{z}}} (0)$$

$$D_2 = \sum_{i=1}^{N_y} \sum_{j=1}^{N_z} \left(\frac{\partial^2 v}{\partial A_i \partial C_j}\right)^{-1} R_{\underline{y}\underline{\hat{z}}} (i - j)$$

The variance of the prediction error can be expressed as

$$E[\underline{\tilde{y}}(k)\underline{\tilde{y}}^T(k)] \quad R[I + A_2 + B_2 + C_2 - 2D_2 - 2E_2]$$

$$= \hat{R}[I - A_2 - B_2 - C_2 + 2D_2 + 2E_2]^{-1}$$

$$\cdot [I + A_2 + B_2 + C_2 - 2D_2 - 2E_2]$$

REFERENCES

1. H. Akaike, "Fitting autoregressive model for prediction," Ann. Inst. Statist. Math. (1969).
2. H. Akaike, "Autoregressive model fitting for control," Ann. Inst. Statist. Math. (1971).
3. H. Akaike, "Statistical predictor identification," Ann. Inst. Statist. Math. (1969).
4. J. Valis, "On-line identification of multivariable linear systems of unknown structure from input-output data," IFAC Symposium on identification and process parameter estimation, (1970).
5. C. W. Chan, C. J. Harris and P. E. Wellstead, "An order-testing criterion for mixed autoregressive-moving-average processes," Int. J. Control, (1974).
6. P. S. Dwyer, "Some applications of matrix derivatives in multivariate analysis," American Stat. Assoc. Journal, (1967).
7. H. Nendecker, "Some theorems on matrix differentiation with special reference to Kronecker matrix products," American Static. Assoc. Journal, (1969).
8. R. L. Kashyap, "Maximum likelihood identification of stochastic linear systems," IEEE Trans. on AC. (1970).

A NEW APPROACH TO BROAD BANDING*

HERBERT J. CARLIN
School of Electrical Engineering
Cornell University
Ithaca, New York 14853

Abstract

A new idea for treating the broad band matching of an arbitrary load to a resistive generator is presented which leads to a simple technique for the design of a lossless two-port equalizer. The method is a numerical one, and only utilizes load impedance data (e.g. experimental) at real frequencies. No model or analytic impedance function for the load is necessary. Nor is the equalizer topology or analytic form of the system transfer function assumed. The arithmetic is well conditioned and the intricacies of gain bandwidth theory are bypassed. A comparison with analytic gain bandwidth theory is given, as well as an example which broadbands an avalanche diode reflection amplifier directly from measured data.

GAIN-BANDWIDTH PROBLEM

The problem we consider is the broad band matching of a load to a resistive generator with a lossless 2-port equalizer. The load impedance is known (say experimentally) at real frequencies over a prescribed frequency band and maximum transducer power gain between generator and load is to be achieved over the band by appropriate design of the equalizer. In the method described here, the final design is obtained by the direct processing of only the real frequency load data. The beautiful gain-bandwidth theories of Fano(1) and Youla(2) require that the load data be first approximated by a model yielding a driving point function which analytically specifies the characteristics of the model over the entire complex frequency plane. This load function is then processed in a complicated fashion to find the theoretic gain bandwidth limits for the approximated model of the load. To realize the equalizer, an analytic form of transfer function (load plus equalizer) must be assumed and the transfer function parameters plus additional all pass zeros adjusted to satisfy the gain-bandwidth restrictions. These restrictions take either the form of a simultaneous set of difficult algebraic equations or a set of integral expressions which must be satisfied simultaneously. Even if the load model is known, the procedure presents great numerical difficulties which in principle are resolvable but in practice become almost intractable, especially if the load includes more than one or two reactors and resistors.

Direct numerical optimization of the elements of the equalizer is also a difficult matter because these elements relate to the system gain in highly non-linear combinations. Furthermore, such direct optimization requires a specific assumption of equalizer topology.

TECHNIQUE FOR BROAD BAND MATCHING USING REAL FREQUENCY LOAD DATA

The technique proposed here is a numerical method which utilizes the real frequency load data $Z_\ell(j\omega) = R_\ell(\omega) + jX_\ell(\omega)$ (or equivalently $Z_\ell^{-1}(j\omega) = Y_\ell(j\omega) = G_\ell(\omega) + jB_\ell(j\omega)$) over the band of interest. No model is assumed and no analytic approximation of the load is required. Furthermore, the topology of the equalizer and the analytic form of the system transfer function are also not assumed. Instead the Thévinin impedance of the equalizer as seen from the load $Z_q(j\omega) = R_q(\omega) + jX_q(\omega)$ (equivalently $Z_q^{-1} =$

---

*The work reported here was supported under NSF grant # 75-10067ENG.

$Y_q(j\omega) = G_q(\omega) + jB_q'(\omega)$) is described in a simple manner in terms of a number of unknown real parameters which once chosen to optimize gain-bandwidth define the equalizer design. The problem is readily formulated by considering the transducer gain $T(\omega)$ of the load-equalizer system.

$$T(\omega) = \frac{\text{Power to Load}}{\text{Power available from generator}} = \frac{4R_\ell(\omega)\, R_q(\omega)}{|Z_\ell(j\omega) + Z_q(j\omega)|^2} \tag{1}$$

with a similar expression in terms of admittances and conductances. The real frequency load data $R_\ell(\omega)$, $X_\ell(\omega)$ is given, the unknown equalizer reactance $X_q(\omega)$ is of course specified by the equalizer resistance $R_q(\omega)$ to within a Foster function by a Hilbert transform, and with these constraints $R_q(\omega)$, $X_q(\omega)$ are to be found to optimize $T(\omega)$ over the prescribed band. Although not a necessary assumption, we will suppose $T(\omega)$ is required to be approximately flat over the band, and we seek to maximize the minimum in-band transducer gain, T_o.

The numerical simplicity of the method described here comes about because of the way in which the unknown equalizer resistance is represented. Let us suppose that $R_q(\omega)$ is to be found as a number of straight line segments, i. e. semi-infinite slopes with positive frequency break points at $0 < \omega_1 < \omega_2 \ldots < \omega_n$. For simplicity, we suppose that $Z_q(\omega)$ is a minimum reactance function whose real part $R_q(\omega)$ is essentially zero above ω_n. Alternately, we might choose to deal with $Y_q(\omega)$ as a minimum susceptance. The choice is easily made by considering the load data.

The ω_k break points can be divided evenly over the band, or chosen with more insight by qualitatively examining the real frequency load data. The choice of ω_n beyond which $R_q(\omega) = 0$ depends on the roll-off desired.

In any case the ω_k are therefore assumed known (they can be adjusted later if desired), and now the equalizer resistance may be specified as a linear combination of the individual total resistive excursions r_k of each of the straight line segments.

$$R_q(\omega) = r_o + \sum_{k=1}^{n} a_k(\omega) r_k \tag{2}$$

Here $r_o = R_q(o)$, and r_k, any positive or negative real number, is the unknown resistive excursion of the k^{th} straight line segment between break points ω_{k-1} and ω_k. Since $R_q(\omega) = 0$, $\omega > \omega_n$.

$$r_o + \sum_{k=1}^{n} r_k = 0 \tag{3}$$

so that with r_o given there are $n - 1$ unknowns, i. e. r_k, $k = 1, \ldots, n-1$.

Clearly,

$$a_k(\omega) = \begin{cases} 1 & , \ \omega_k \leq \omega \\ \dfrac{\omega - \omega_{k-1}}{\omega_k - \omega_{k-1}} & , \ \omega_{k-1} \leq \omega \leq \omega_k \\ 0 & , \ \omega \leq \omega_{k-1} \end{cases} \tag{4}$$

and provided the r_k are real and $R_q(\omega) \geq 0$, the impedance $Z_q(\omega)$ is passively realizable.

As a result of this representation for $R_q(\omega)$, the (minimum) reactance

$X_q(\omega)$ is also specified as a linear combination of the same unknown resistive excursions, the r_k. Thus

$$X_q(\omega) = \sum_{k=1}^{n} b_k(\omega) r_k \tag{5}$$

and for any frequency ω the $b_k(\omega)$ are immediately known in terms of the break point frequencies and are of course independent of the r_k.

$$b_k(\omega) = \frac{1}{(\omega_k - \omega_{k-1})\pi} \int_{\omega_{k-1}}^{\omega_k} \ell n \left| \frac{y+\omega}{y-\omega} \right| dy \tag{6}$$

The integral (6) has a simple closed form evaluation$^{(3)}$, which need not be repeated here.

Once the line segments describing $R_q(\omega)$ have been found, an appropriate rational approximation $\hat{R}_q(\omega) \cong R_q(\omega)$ is determined by any convenient means (least squares used here) and $\hat{Z}_q(j\omega)$ realized as a Darlington reactance 2-port (the equalizer) with a resistive termination (the generator impedance). Sophisticated numerical procedures exploiting (1), (2), (5), (6) can of course be formulated, but because of the form of gain equation (1), and the representation of $R_q(\omega)$ and $X_q(\omega)$ as linear combinations of the real unknowns r_k, the gain is at most quadratically dependent on the unknowns and, for example, a Newton-Raphson optimization routine can be set up with a Jacobian which is a linear function of the unknowns.

Rather than employ an elaborate optimization routine, a straightforward iterative procedure was used in order to illustrate the simplicity of the basic idea. Thus for a low-pass problem the d. c. gain level is set and the associated r_o calculated. The initial guess for the resistance excursions r_k is chosen to sustain the d. c. gain at the break points assuming reactance cancellation. The actual reactances corresponding to the resulting line segments are then calculated and the r_k readjusted with the reactance taken into account and the minimum gain over the band noted. The d. c. gain is then increased and the procedure repeated until the minimum gain can no longer be increased.

The following observations indicate why this straightforward numerical procedure is so well behaved. 1. The gain at any frequency is weakly affected by the contributions of line segments remote from the frequency in question. 2. When the total reactance $X_T(\omega) = X_\ell(\omega) + X_q(\omega)$ is less than about $0.3\,(R_\ell(\omega) + R_q(\omega))$, its effect on $T(\omega)$ is small. 3. Given $X_T(\omega)$, maximum gain occurs at ω when

$$R_q(\omega) = \sqrt{R_\ell^2(\omega) + X_T^2(\omega)} \quad ,$$

4. The arithmetic is very simple and the gain level depends directly on the unknown r_k.

A low pass example was chosen in which the real frequency load data was generated by an L-C-R impedance* whose equalization has been worked out by gain-bandwidth theory in Fano's paper.$^{(1)}$ He shows that for a load with input series L = 2.3 h., shunt C = 1.2 fd., R = 1 ohm, the idealized absolutely flat gain over unit band $0 \leq \omega \leq 1$ is $T_{0I} = .924$. He also obtains the parameters of a realizable Chebychev equalizer transfer function containing an all pass zero which yields a matching network with 4 energy storage elements including a pair of coupled coils. This gives an equal ripple pass band gain flat to ± .083 dB, with a minimum of $\hat{T}_0 = 0.816$.

---

*This problem is not simple by gain bandwidth theory, but a theoretical solution is known and hence comparisons are possible.

In employing the present procedure it was decided to use 5 slopes, i.e. four unknowns because of (3). (Note that a relatively few straight line segments may be sufficient to describe even a relatively high-order equalizer or conversely). Since the load reactance is highly inductive with increasing frequency, the equalizer impedance $Z_q(\omega)$ was chosen as minimum reactance to permit a final shunt capacitor in the equalizer. The break points were taken as $\omega_1 = .25$, $\omega_2 = .5$, $\omega_3 = .75$, $\omega_4 = 1.0$, $\omega_5 = 1.25$. There are two essentially different solutions depending on whether $r_o > R_\ell(0)$, or $r_o < R_\ell(0)$. Fano's Chebychev network is for $r_o < R_\ell(0)$.

The two d.c. values of $r_o = R_q(0)$ are related to the d.c. gain T(0) by

$$r_o = R_\ell(0)\,[k_o \pm \sqrt{k_o^2 - 1}], \quad k_o = \frac{2}{T(0)} - 1 \tag{7}$$

For the L-C-R problem considered the calculated line segment resistance characteristics are shown on Fig. 1. The two straight line solutions for $r_o > R_\ell(0)$ and $r_o < R_\ell(0)$ are given as $R_{aq}(\omega)$, $R_{bq}(\omega)$ respectively. For the former case $T_0 = 0.846$, for the latter case $T_0 = 0.832$. The resistance characteristic for the case $r_o > R_\ell(0)$ was then approximated by a resistance function realizable as a ladder, i.e. the inverse of an even positive polynomial, and is shown as $\hat{R}_{aq}(\omega)$ on Fig. 1. The finally realized gain curve corresponding to $\hat{R}_{aq}(\omega)$ is plotted on Fig. 2. The minimum gain for the actual equalizer is $\hat{T}_0 = 0.837$. The gain is flat to within ± .12 dB, and the circuit as shown on Fig. 2, is simply a three-element low-pass ladder. If the value of the final break-point is increased to $\omega_3 = 1.5$ so as to permit gentler roll-off, the process yields a 2 element ladder equalizer with minimum gain $\hat{T}_0 = 0.81$. If ω_n is chosen to approach 1, the gain approaches the ideal flat gain optimum.

A rational approximation to the resistance characteristic (b) of Fig. 1 was also carried out for the case $r_o < R_\ell(0)$, and is shown on the figure as $\hat{R}_{bq}(\omega)$. The resultant structure (not shown) has five energy storage elements including a pair of coupled coils (as in Fano's solution) with minimum gain over the band of $\hat{T}_0 = 0.827$. This is slightly higher than Fano's result but one additional energy storage element is used. The gain is flat to within ± .086 dB, $0 \leq \omega \leq 1$.

The minimum gain and ripple in our example compare closely with Fano's results which is quite remarkable considering that his structure corresponds to an analytically optimized equal ripple gain function whereas only a simple numerical procedure was used here. A summary of the results is given in the table on the following page.

A second example (band-pass instead of low-pass) based directly on experimental data is the broad banding of a microwave reflection amplifier whose active element, an avalanche diode, is connected to a 3-port circulator through a lossless equalizer. The transducer gain in this case is

$$G(\omega^2) = \frac{(R_q(\omega) - R_\ell(\omega))^2 + (X_q(\omega) + X_\ell(\omega))^2}{(R_q(\omega) + R_\ell(\omega))^2 + (X_q(\omega) + X_\ell(\omega))^2} \tag{8}$$

with $G > 1$ when $R_\ell < 0$.

Table: Comparison of Real Frequency and Analytic Theory (LCR Load)

	r_o	# of Elem.	# of Transformers	Min.** Gain		Loss Ripple (dB)
				T_o	$\hat{T}_o$	
Real Freq.	$r_{o1} > R_o$	3	0	0.846	0.837	± 0.12
	$r_{o2} < R_o$	4	1	0.820	0.79	± 0.2
	$r_{o2} < R_o$	5	1	0.832	0.827	± 0.086
* Anal. Theo.	$r_{o2} < R_o$	4	1	----	0.816	± 0.083

* Chebychev optimization by Fano(1).

** Ideal flat gain $T_{0I} = 0.924$

The measured data of the diode in its mount[†] is shown in Fig. 3. The straight line conductance characteristic was determined to approximate flat gain over a 15 percent frequency band and is shown in Fig. 4 together with the rational approximation to $G_g(\omega)$. The resulting gain curve and equalizer are given in Fig. 5. The gain is 8.53 ± 0.13 dB over a 15 percent band (4.8 to 5.5 GHz), and 8.23 ± 0.45 dB over a 20 percent band. It will be noted that the equalizer matches this heavily loaded low impedance diode directly into the 50 ohm line without any additional impedance transformer. Furthermore, real frequency stability criteria are readily incorporated in the procedure.

Acknowledgement:

The assistance of Jim Komiak of Cornell University in performing the computations is gratefully acknowledged.

References:

1. F. M. Fano, "Theoretical Limitations on the Broadband Matching of Arbitrary Impedances", J. Franklin Inst., Vol. 249, pp. 57-83, January 1960 and pp. 139-155, February 1960.

2. D. C. Youla, "A New Theory of Broadband Matching", IEEE Trans. of Circuit Theory, Vol. CT-11, No. 1, pp. 30-50, March 1964.

3. H. W. Bode, "Network Analysis and Feedback Amplifier Design", D. Van Nostrand 1945, see p. 319.

[†] The data was furnished by courtesy of J. Berenz, G. C. Dalman and C. A. Lee of Cornell University.

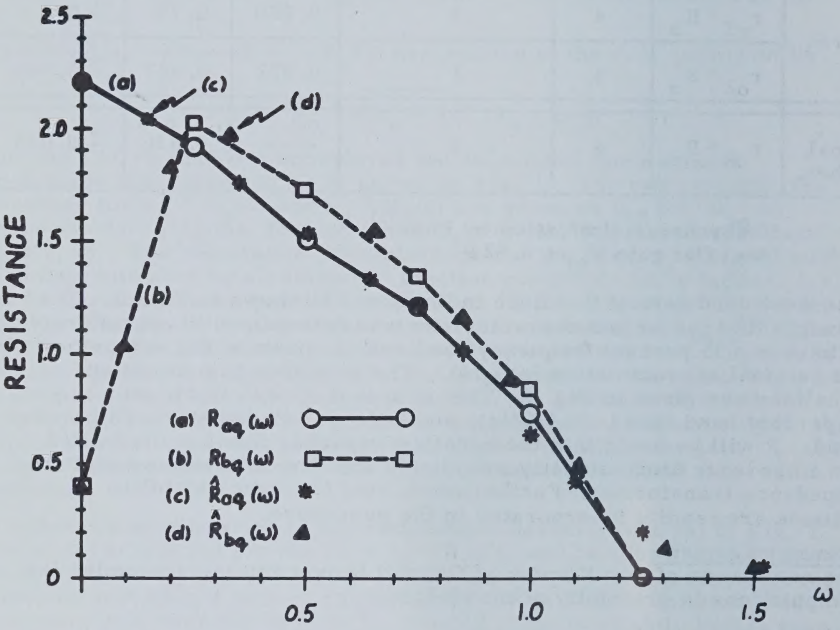


Figure 1. **Resistance characteristics of equalizer for L-C-R load.**

a) $R_{aq}(\omega)$, **straight line characteristic,** $r_o > R_\ell(0)$, $T_o = 0.846$

b) $R_{bq}(\omega)$, **straight line characteristic,** $r_o < R_\ell(0)$, $T_o = 0.832$

c) $\hat{R}_{aq}(\omega)$, **Rational approximation to** $R_{aq}(\omega)$, $\hat{T}_o = 0.837$

$$\hat{R}_{aq}(\omega) = \frac{2.2}{1 + 2.56\omega^2 - 4.44\omega^4 + 4.29\omega^6}$$

d) $\hat{R}_{bq}(\omega)$, **Rational approximation to** $R_{bq}(\omega)$, $\hat{T}_o = 0.827$

$$\hat{R}_{bq}(\omega) = \frac{0.42 + 89.7\omega^2}{1 + 25.06\omega^2 + 153.7\omega^4 - 273.80\omega^6 + 231.54\omega^8}$$

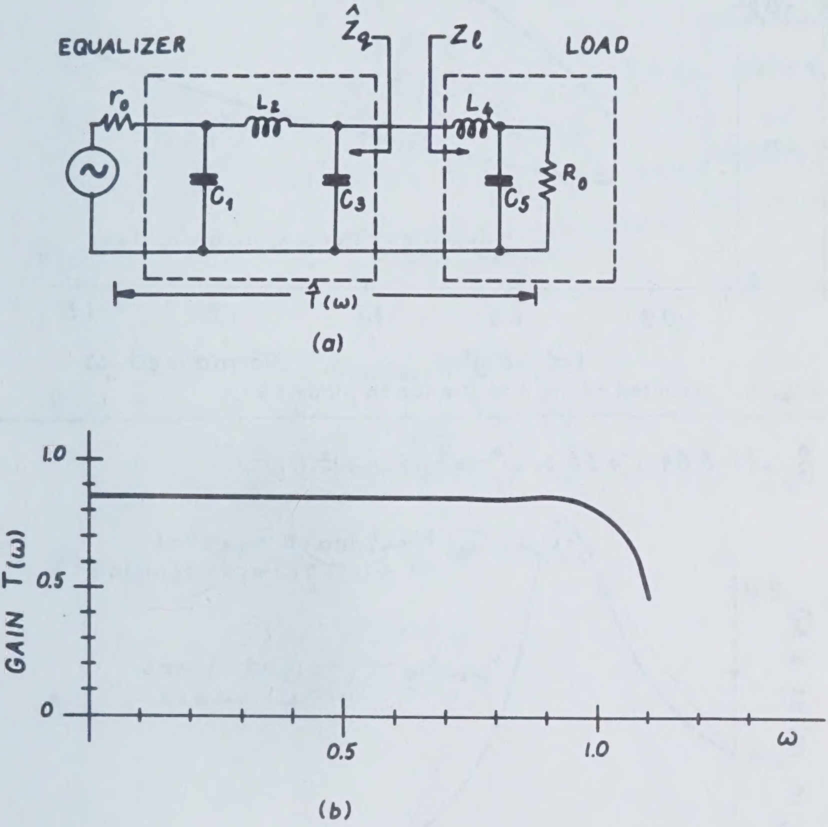
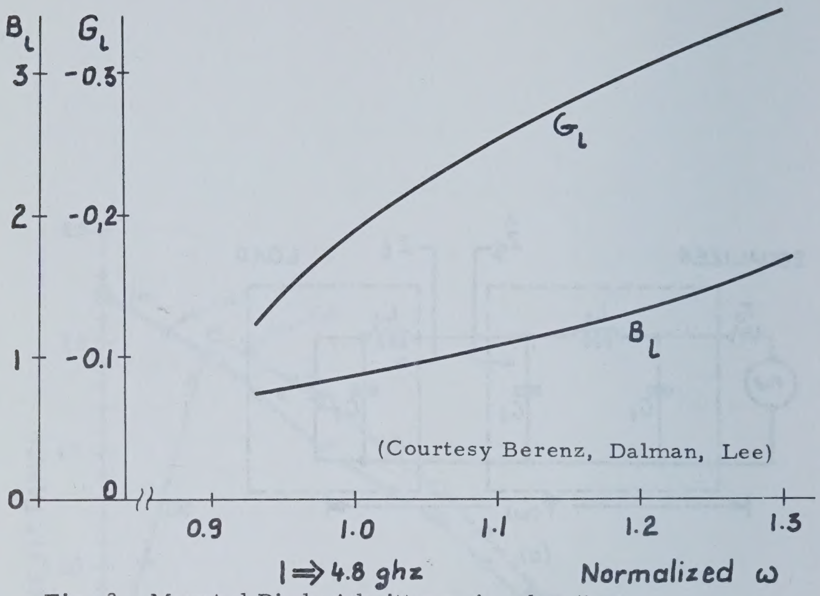

EQUALIZER
$\hat{Z}_q$
Z_ℓ
LOAD
r_0
L_2
L_4
C_1
C_3
C_5
R_0
$\hat{T}(\omega)$
(a)
1.0
0.5
0
GAIN $\hat{T}(\omega)$
0.5
1.0
ω
(b)

Figure 2. Equalizer realization for L-C-R load, $r_o > R_\ell(0)$

a) Equalizer: $r_o = 2.2$, $C_1 = 0.351$, $L_2 = 2.89$, $C_3 = 0.932$

Load: $L_4 = 2.3$, $C_5 = 1.2$, $R_o = 1.0$

b) Transducer gain $\hat{T}(\omega)$ vs ω flat to $\pm$ 0.12 dB, $\hat{T}_o = 0.837$.



(Courtesy Berenz, Dalman, Lee)

Fig. 3. Mounted Diode Admittance in mhos × 50.

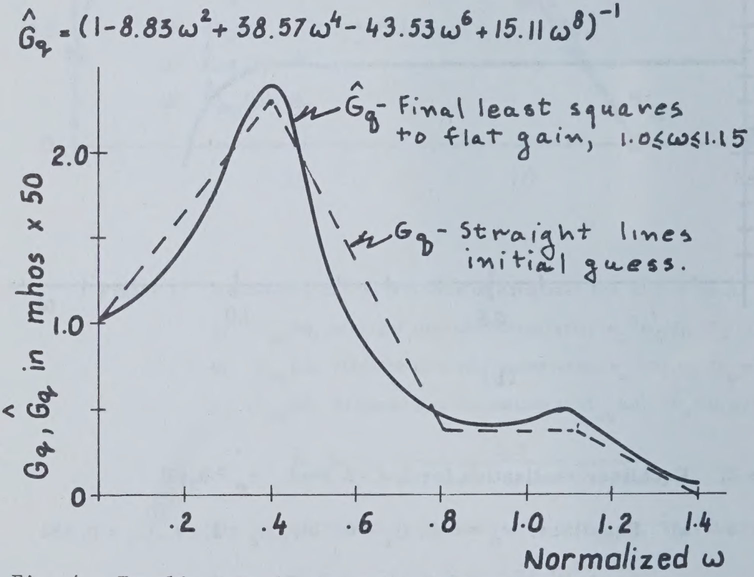


Fig. 4. Equalizer Conductance Characteristics.

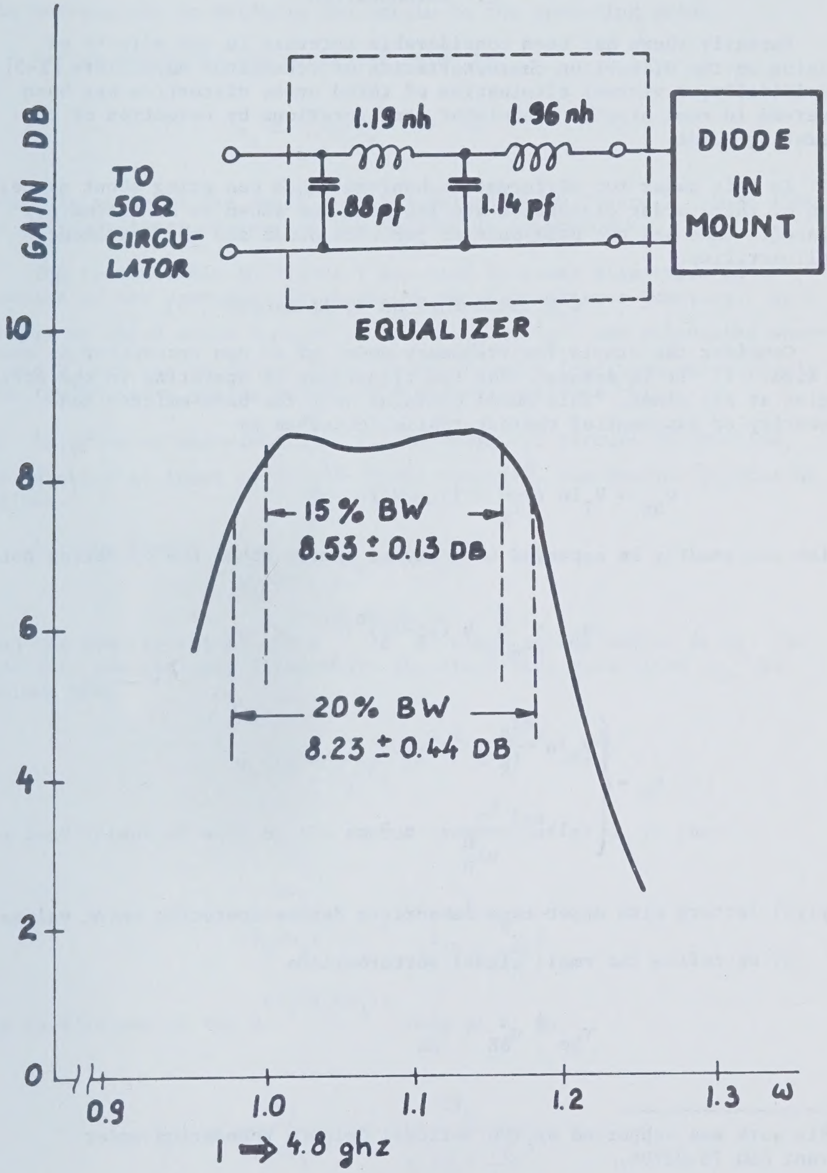


Fig. 5. Avalanche Diode Amplifier

SMALL-SIGNAL THIRD-ORDER DISTORTION ANALYSIS OF TRANSISTOR AMPLIFIERS[†]

J. D. COBB AND T. N. TRICK
University of Illinois
Urbana, Illinois 61801

I. INTRODUCTION

Recently there has been considerable interest in the effects of biasing on the distortion characteristics of transistor amplifiers [1-3]. Specifically, a virtual elimination of third order distortion has been observed in some bipolar transistor configurations by selection of the operating point.

In this paper two different mechanisms which can bring about cancellation of third order distortion are isolated and shown to be intimately related. They are the base-emitter junction diode and the avalanche nonlinearities.

II. BASE-EMITTER NONLINEARITY

Consider the simple low-frequency model of an npn transistor as shown in figure 1. It is assumed that the transistor is operating in the active region at all times. This model contains only the base-emitter nonlinearity or exponential characteristic described by

$$v_{BE} = V_T \ln \left(\frac{\beta i_B}{I_S} + 1\right) \simeq V_T \ln \frac{\beta i_B}{I_S} \tag{1}$$

which can readily be expanded in a Taylor series about the operating point.

$$v_{BE} = \sum_{n=o}^{\infty} b_n (i_B - I_B)^n \tag{2}$$

$$b_n = \begin{cases} V_T \ln \dfrac{\beta I_B}{I_S}, & n = o \\ (-1)^{n+1} \dfrac{V_T}{n I_B^n}, & n > o \end{cases} \tag{3}$$

Capital letters with upper case subscripts denote operating point values.

If we define the small signal perturbations

$$v_{be} \triangleq v_{BE} - V_{BE} \tag{4}$$

[†] This work was supported by the National Science Foundation under Grant ENG 75-02708.

$$i_b \triangleq i_B - I_B \tag{5}$$

then we can write

$$v_{be} = \sum_{n=1}^{\infty} b_n i_b^n \quad . \tag{6}$$

This corresponds to shifting the origin to the operating point.

Let us represent the nonlinear terms by

$$v_e \triangleq \sum_{n=2}^{\infty} b_n i_b^n \tag{7}$$

then this suggests the small signal model shown in figure 2. Note that the only nonlinear element is the distortion generator v_e.

The test circuit in figure 3 was used to study distortion as a function of the operating point due to the base-emitter junction. As a measure of third order distortion the quantity $\left|\frac{H_3}{H_1}\right|$ was calculated where H_3 and H_1 are the third and first order transfer functions respectively from the Volterra series method of nonlinear analysis [4].

In order to calculate $\left|\frac{H_3}{H_1}\right|$ for the specific circuit in question, the exponential input method was used. Briefly, the derivation goes as follows.

Let the input

$$v_s(t) = e^{S_1 t} + e^{S_2 t} + e^{S_3 t} \quad . \tag{8}$$

Then the coefficient of the $e^{(S_1+S_2+S_3)t}$ term at the output is H_3. To find this coefficient, first short the distortion generation v_e. It follows that

$$i_b(t) = \frac{1}{R_S+b_1}\left(e^{S_1 t} + e^{S_2 t} + e^{S_3 t}\right) \quad . \tag{9}$$

The coefficient of each of the second order modes of v_e is then

$$\frac{2b_2}{(R_S+b_1)^2} = -\frac{V_T}{I_B^2\left(R_S+\frac{V_T}{I_B}\right)^2} \quad . \tag{10}$$

The coefficient of the $e^{(S_1+S_2+S_3)t}$ mode of v_e is

$$\frac{6b_3}{(R_S+b_1)^3} = \frac{2V_T}{I_B^3\left(R_S+\frac{V_T}{I_B}\right)^3} \quad . \tag{11}$$

Next, find the coefficient of the third order mode v_e due to the interaction of second order and first order terms. The coefficient of each of the second order terms of i_b is

$$-\frac{2b_2}{(R_S+b_1)^3} = \frac{V_T}{I_B^2(R_S+\frac{V_T}{I_B})^3} \quad . \tag{12}$$

The third order mode of v_e due to this interaction then has the coefficient

$$6b_2\left[-\frac{2b_2}{(R_S+b_1)^3}\right]\left(\frac{1}{R_S+b_1}\right) = -\frac{3V_T^2}{I_B^4(R_S+\frac{V_T}{I_B})^4} \quad . \tag{13}$$

Finally, short the input generator v_S and find the coefficient of the third order term at the output.

$$H_3 = \frac{\beta I_B V_T R_L(2I_B R_S - V_T)}{(I_B R_S + V_T)^5} \tag{14}$$

By inspection

$$H_1 = -\frac{\beta I_B R_L}{(I_B R_S + V_T)} \quad . \tag{15}$$

Therefore

$$\left|\frac{H_3}{H_1}\right| = \frac{V_T|2I_B R_S - V_T|}{(I_B R_S + V_T)^4} \quad . \tag{16}$$

If

$$v_s(t) = A \cos wt \tag{17}$$

then there are contributions to the cos 3 wt mode at the output due to all third and higher odd order transfer functions. If A is small enough then higher order contributions can be neglected and the third order component of the output voltage is

$$\frac{1}{2^2}\left(\frac{1}{3!}\right) A^3 H_3 = \frac{A^3}{24} H_3 \quad . \tag{18}$$

If we define $v_o^{(3)}$ and $v_o^{(1)}$ as the third harmonic and fundamental rms amplitudes of the output voltage, then

$$\frac{v_o^{(3)}}{v_o^{(1)}} = \left|\frac{\frac{A^3}{24} H_3}{AH_1}\right| = \frac{A^2}{24} \left|\frac{H_3}{H_1}\right| . \tag{19}$$

Clearly, for constant input level, $\left|\frac{H_3}{H_1}\right|$ is approximately proportional to $\frac{v_o^{(3)}}{v_o^{(1)}}$ and is therefore a valid measure of normalized third order distortion.

If we set

$$I_B = \frac{V_T}{2R_S} \tag{20}$$

then

$$\left|\frac{H_3}{H_1}\right| = 0 \tag{21}$$

and third order distortion is virtually eliminated. The ratio $\left|\frac{H_3}{H_1}\right|$ is plotted in figure 4.

Measurements were made in the laboratory which seem to verify these results. Data was taken for several values of load resistance. All readings were taken with A = 28.3 mV. Figure 5 shows the relationship between measured and theoretical values. Note the characteristic drop in distortion near the collector current corresponding to a base current of $\frac{V_T}{2R_S}$.

In figure 6 a similar set of curves is plotted, this time for a larger source resistance. Again a striking reduction in distortion can be observed near the predicted value of collector current.

The slight inconsistencies between predicted and measured values are probably due to problems in the experimental set up. Predicted values were calculated assuming an operating temperature of 25°C while the actual temperature of operation may have been somewhat higher, accounting in part for the obvious shift in the position of the point of minimum distortion. Also, some distortion was introduced by the source generator.

It should be noted that the actual levels of distortion are very close to those predicted. This indicates the well-known dominance of the base-emitter nonlinearity in the active region. However, the obvious conclusion, as far as third order distortion is concerned, is that this dominance can be eliminated by correctly biasing the base-emitter junction, leaving open the possibility of significant interactions between two or more nonlinear effects. These ideas serve as motivation for the next section.

III. AVALANCHE AND BASE-EMITTER INTERACTIONS

Consider the transistor model shown in figure 7. Again the base-emitter nonlinearity is included. In this model, however, the dependent current source takes into account the avalanche breakdown effect of the collector-base junction which is governed by the equation

$$i_C = \frac{\beta i_B}{1 - (\frac{V_{CB}}{V_{CBO}})^4} \quad . \tag{22}$$

V_{CBO} is the collector-base breakdown voltage of the transistor.

This expression can be expanded in a Taylor series about the operating point

$$i_C = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} a_{nm}(i_B - I_B)^n \, (v_{CB} - V_{CB})^m \quad . \tag{23}$$

It is convenient that

$$a_{nm} = 0 \; , \quad n \geq 2 \quad . \tag{24}$$

If we define

$$v_{cb} \stackrel{\Delta}{=} v_{CB} - V_{CB} \tag{25}$$

$$i_c \stackrel{\Delta}{=} i_C - I_C \tag{26}$$

then

$$i_c = \sum_{m=1}^{\infty} a_{om} v_{cb}^m + \sum_{m=0}^{\infty} a_{1m} i_b v_{cb}^m \quad . \tag{27}$$

Letting

$$i_e \stackrel{\Delta}{=} \sum_{m=2}^{\infty} a_{om} v_{cb}^m + \sum_{m=1}^{\infty} a_{1m} i_b v_{cb}^m \tag{28}$$

suggests the small signal model of figure 8.

Becuase of the complexity of even a simple model such as this, a digital computer was used to perform the analysis. Figure 9 shows a typical pair of curves generated by adjusting collector-base voltage while holding base current constant. Null points can be observed for various values of bias voltage. The breakdown voltage of the transistor is approximately 40 v.

In general there seems to be two ranges of base current at which these nulls occur. The position of the null is plotted versus base current in figure 10.

At first it may seem rather strange that there can be a complete cancellation between base-emitter distortion and the normally less significant avalanche nonlinearity. However, the reason for this sort of interaction can be readily seen if one observes that the regions where nulls exist correspond exactly to those regions where distortion due to the base-emitter nonlinearity alone is small. Thus, if the transistor is biased with an appropriate base current, the third order distortion due to the base-emitter nonlinearity alone is small enough for a null to exist at some collector-base voltage. The complete cancellation is due to interaction of the two nonlinear effects.

It is interesting to note that the slope of the curve in figure 10 is always negative. This indicates that the collector-base voltage at which the null occurs can be decreased by increasing the base current.

These results seem to be true in general. The position of the null can be placed almost anywhere by adjusting the base current. The two regions in which nulls occur can be positioned by adjusting the various parameters in the circuit. They can even be made to overlap giving two nulls at the same base current.

IV. CONCLUSION

It has been shown that third order distortion can be reduced by selecting an appropriate operating point. First, the base-emitter junction should be biased to bring down the level of distortion due to the base-emitter nonlinearity alone. Then, the collector-base voltage can be adjusted to bring about a further reduction in distortion due to the interaction of the two nonlinear effects.

V. REFERENCES

[1] J. Eachus, "Distortion in Ultralinear Solid-State Devices," IEEE Journal of Solid-State Circuits, Vol. SC-10, pp. 485-497, Dec. 1975.

[2] L. C. Thomas, "Eliminating Broadband Distortion in Transistor Amplifiers," Bell System Technical Journal, March 1968.

[3] A. J. Mallinckrodt, F. M. Gardner, "Distortion in Transistor Amplifiers," IEEE Transactions on Electron Devices, pp. 288-289, July 1963.

[4] S. Narayanan, "Transistor Distortion Analysis Using Volterra Series Representation," Bell System Technical Journal, Vol. 46, pp. 991-1023, May-June 1967.

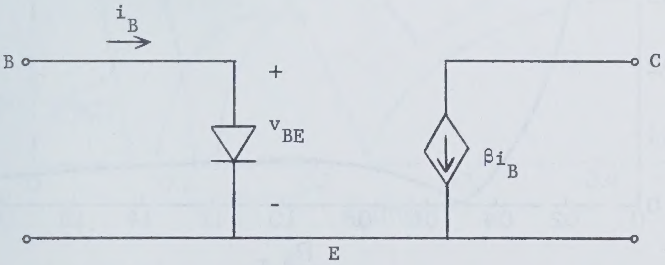


Figure 1. Low frequency model with base-emitter nonlinearity only.

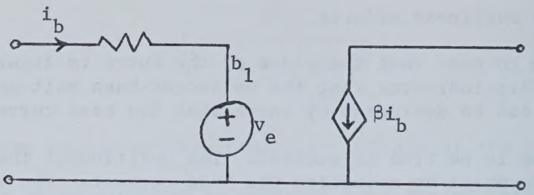

i_b
b_1
+
−
v_e
βi_b

Figure 2. Small signal model.

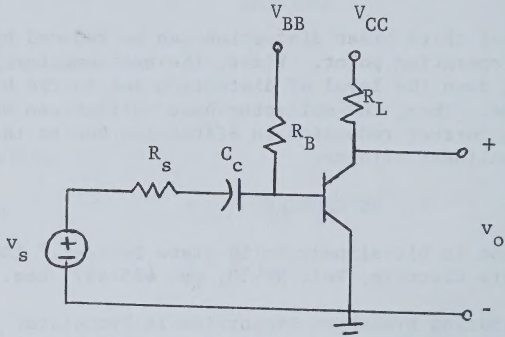

V_{BB}
V_{CC}
R_L
R_B
R_s
C_c
+
v_o
v_s
−

Figure 3. Test circuit.

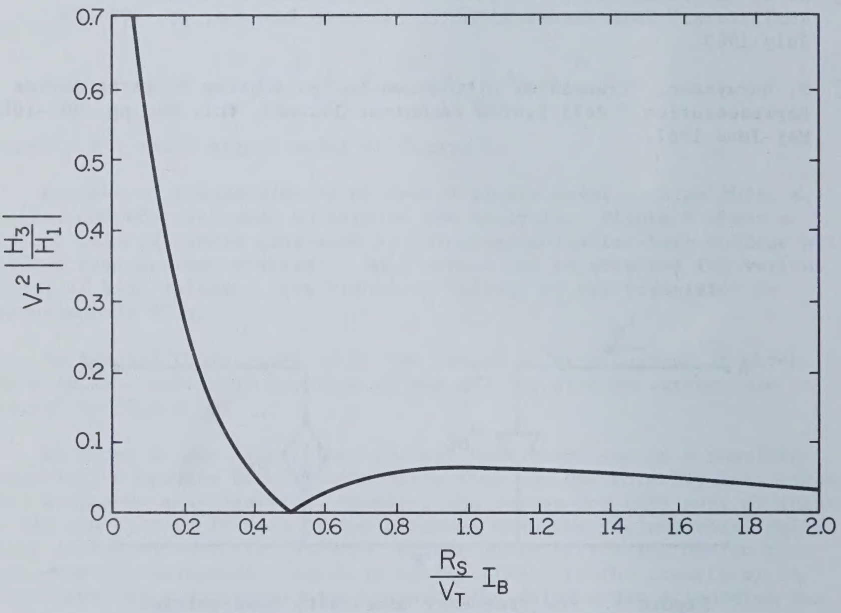


Figure 4. Third order distortion versus base current.

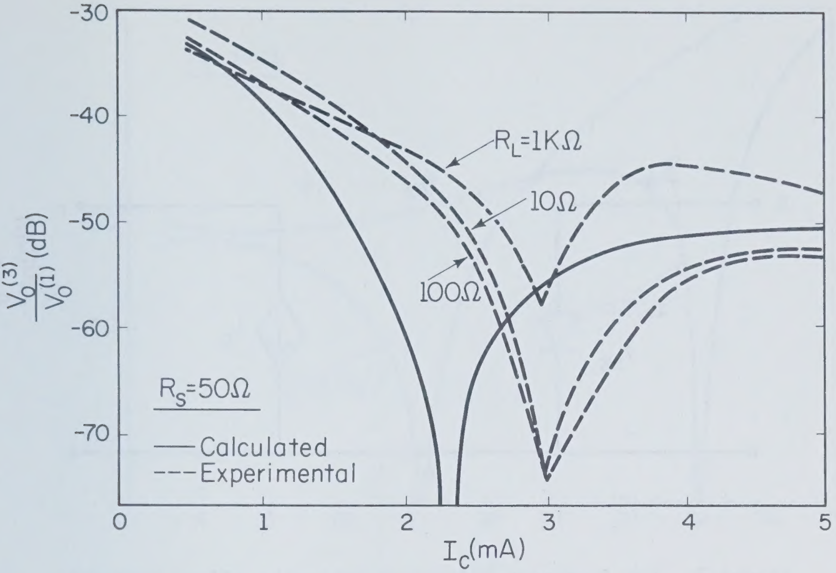


Figure 5. Measured and theoretical distortion for $R_s = 50\Omega$.

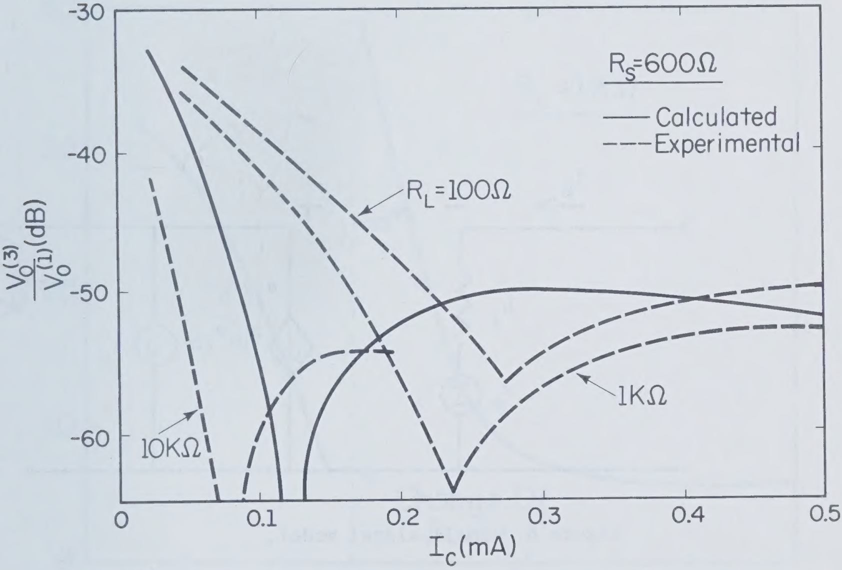


Figure 6. Measured and theoretical distortion for $R_s = 600\Omega$.

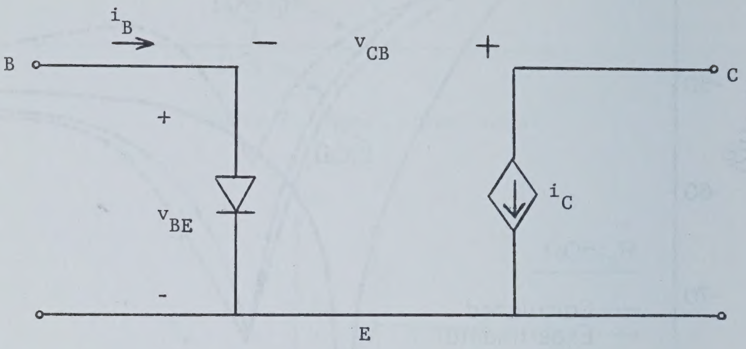

i_B
−
v_CB
+
B
C
+
v_BE
i_C
-
E

Figure 7. Low frequency model with avalanche effect included.

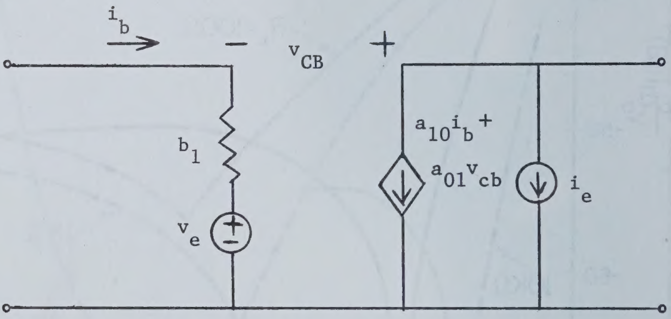

i_b
−
v_CB
+
b_1
v_e
a_10 i_b +
a_01 v_cb
i_e

Figure 8. Small signal model.

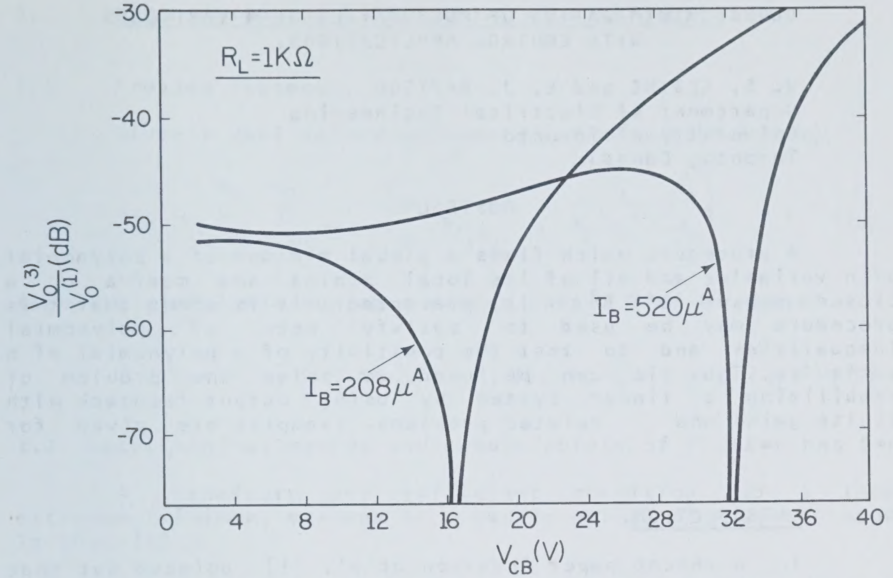


Figure 9. Third order distortion versus collector-base voltage.

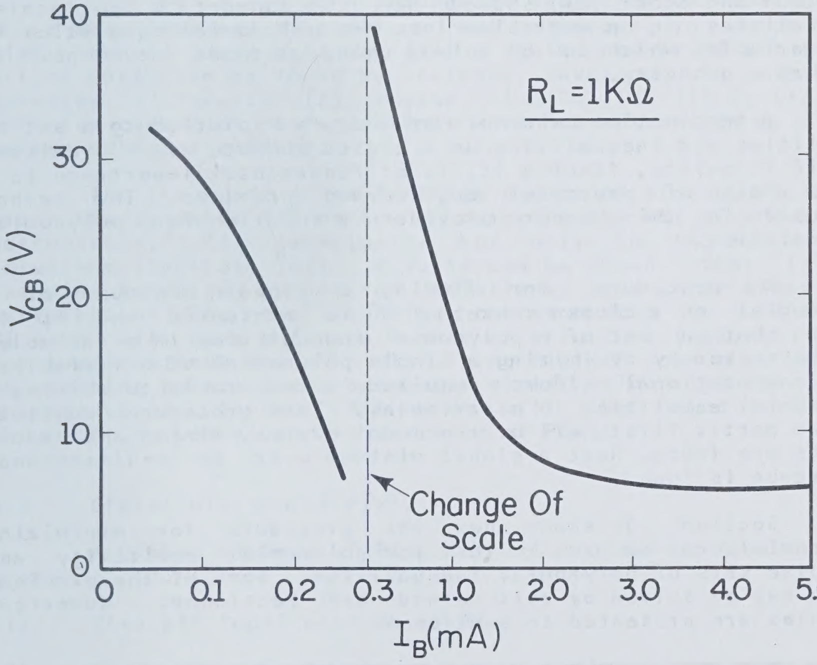


Figure 10. Position of null versus base current.

GLOBAL MINIMIZATION OF POLYNOMIALS OF N VARIABLES WITH CONTROL APPLICATIONS.

W. S. GESING and E. J. DAVISON
Department of Electrical Engineering
University of Toronto
Toronto, Canada.

ABSTRACT.

A procedure which finds a global minimum of a polynomial of n variables and all of its local minima and maxima on a closed subset of R^n is presented. It is shown that this procedure may be used to satisfy sets of polynomial inequalities and to test the positivity of a polynomial of n variables. Thus it can be used to solve the problem of stabilizing a linear system by using output feedback with finite gains and related problems. Examples are given for n=2 and n=3.

1. INTRODUCTION.

In a recent paper Anderson et al. [1] pointed out that several problems in control theory may be formulated as a set of multivariable polynomial inequalities and that these inequalities in principle can be solved in a finite number of steps using methods of decision theory. In a later paper Anderson and Scott [2] showed how to convert m polynomial inequalities in n variables into n+m polynomial equalities in n+m variables which can be solved using methods developed in algebraic geometry.

Determination whether or not a solution to a set of equalities and inequalities on a closed subset of R^n exists, and if it exists, finding it, is of fundamental importance to a large class of practical engineering problems. The method proposed in this paper provides a solution for polynomial inequalities.

A procedure for finding a global minimum of a polynomial on a closed subset of R^n is presented and it is shown that a set of m polynomial inequalities in n variables may be solved by minimizing a single polynomial of n variables. The computational effort required is comparable to solving n polynomial equalities in n variables. This procedure consists of two parts. First, all of the local maxima, minima and saddle points are found. Next a global minimum over an n-dimensional hypercube is found.

Section 3 shows how the procedure for minimizing polynomials can be used to test for polynomial positivity and to solve sets of polynomial inequalities. Some of the problems which may be solved by this method are mentioned. Numerical examples are presented in section 4.

This work has been supported by the National Research Council of Canada under Grant A4396.

2. MINIMIZATION OF A POLYNOMIAL OF N VARIABLES.

2.1 Problem statment.

Given a real valued polynomial of n variables P(x)

$$P(x) = \sum_{i_1=0}^{q_1} \sum_{i_2=0}^{q_2} \cdots \sum_{i_n=0}^{q_n} a_{i_1 i_2 \ldots i_n} x_1^{i_1} x_2^{i_2} \ldots x_n^{i_n} \qquad (1a)$$

find its global minimum on a closed subset H of R^n where

$$H = \{x \in \mathbb{R}^n | x_i^- \le x_i \le x_i^+ ,\ i = 1, 2, \ldots n,\ x_i^- \in \mathbb{R} ,\ x_i^+ \in \mathbb{R} \} \qquad (1b)$$

2.2 Local minima, maxima and saddle points of P(x).

A necessary and sufficient condition for a local extremum (minimum, maximum or a saddle point) of P(x) to occur is that [3] :

$$\frac{\partial P(x)}{\partial x_i} = 0 \qquad i = 1, 2, \ldots n \qquad (2)$$

(2) is a set of n polynomial equations in n variables. These equations may have (i) no real solutions, in which case P(x) has no local extrema; (ii) a finite number of solutions, all of which can be found by methods developed in algebraic geometry (Anderson [2] , Moses [4] , Collins [5]); (iii) an infinite number of real solutions which correspond to an infinite number of local extrema.

In the case of an infinite number of minima, it is not clear how to solve (2) or how to describe the set of solutions. Fortunately, (2) generically has only $(q_1 + q_2 + \ldots + q_n - 1)^n$ solutions (Coolidge [6]). Also it can be shown that if (2) has an infinite number of solutions, then, for almost any perturbation ΔP, the equation corresponding to minimizing $P(x) + \Delta P$ will have a finite number of solutions (Anderson [2]). Any local extrema found in minimizing $P(x) + \Delta P$ can then be used as starting points for a parameter minimization algorithm (e.g. Rosenbrock [7], Davison [8]) which will then find an exact value of the corresponding extremum of P(x).

2.3 Global minimum of P(x).

A global minimum of P(x) for all x∈H can be found in two steps:

(1) Find all local extrema of P(x).

(2) Find all the local extrema of P(x) over each of the 2n (n-1)-dimensional hypercubes forming the boundaries of H.

The global minimum of P(x) over H is then the smallest value of P(x) evaluated at local extrema found either in step (1) or step (2) and located either inside H or on one of its boundaries.

In actual computation, the number of minimizations is smaller than that indicated by the algorithm above since some of the "edges" and "corners" of the hypercube are shared by two or more of its "faces". For example, for n=3 H becomes a cube, and step (1) involves solving 3 polynomial equations in 3 variables, step (2) consists of 6 minimizations over each of the 2 dimensional faces of H (involving the solution of 2 polynomial equations in 2 variables each), 12 minimizations over the 12 edges of the cube (1 polynomial equation in 1 variable) and finally evaluating P(x) at the eight corners of H.

3. APPLICATION TO PROBLEMS IN CONTROL THEORY.

3.1 Polynomial positivity test.

Bose and Modarressi [9] presented a procedure for a global positivity test for a multivariable polynomial by computing inner determinants of polynomial matrices, finding zeros of the resulting polynomial equations and then testing "joint intervals".

The procedure for finding a global minimum of a multivariable polynomial presented here can be used to test the positivity of polynomials over an arbitrary closed hypercube in R^n by simply finding a global minimum of a polynomial over that hypercube. If the global minimum is positive, then so is the polynomial which is tested. It should be noted that polynomials of several variables are extremely sensitive to infinitesimally small perturbations in their coefficients due to cancellation of leading coefficients along some curves in R^n. It is therfore unlikely that any test which relies on numerical computation of determinants can be relied on to test for a global positivity of polynomials for all $x \in R^n$. However restriction of x to a closed subset of R^n removes this difficulty.

3.2 Solution of a set of polynomial inequalities.

The following problem is considered:

Problem:

Determine if there exists $x \in H$ (defined by (1b)) such that m polynomial inequalities in n variables

$$p_i(x) > 0 \quad i = 1, 2, \ldots m \tag{3}$$

are satisfied. If a solution exists, find at least one $x \in H$ which satisfies (3).

Define

$$P(x) \triangleq \prod_{i=1}^{m} p_i(x) \qquad (4)$$

Here $P(x)$ is a single polynomial in n variables which is a continuous function of x. $P(x)$ is equal to zero if and only if one or more of $p_i(x)$ are equal to zero, and it is positive if all the $p_i(x)$ are positive or if some of the $p_i(x)$ are pairwise negative. These properties of $P(x)$ suggest the following theorem:

Theorem 1.

The set of m polynomial inequalities in n variables (3) has a solution if and only if at least one of the local maxima of $P(x)$ on H, or on the boundaries of H satisfies (3).

Proof:

Due to page restrictions all proofs are omitted; they can be found in [13].

Solution:

To find a solution to a set of inequalities (3) on H construct $P(x)$ and find all of its local maxima on H and its boundaries. Check if any of the maxima satisfy (3). If not, then there exist no $x \in H$ which satisfy (3).

4. EXAMPLES.

4.1 Example 1: Local extrema of a polynomial.

It is desired to find all of the local extrema of the following polynomial of 2 variables

$$P(x,y) = x^2y^4 - 2xy^3 + (1+2x^2)y^2 + (6x-2x^2)y + (50+x^2-2x^3+x^4) \qquad (5)$$

In this case one has to find all solutions of the following system of two polynomial equations in two unknowns:

$$2xy^4 - 2y^3 + 4xy^2 + (6-4x)y + (2x-6x^2+4x^3) = 0 \qquad (6)$$

$$4x^2y^3 - 6xy^2 + (2+4x^2)y + 6x - 2x^2 = 0$$

The necessary and sufficient conditions for the equations (6) to have common roots is that their resultants vanish [10 p298], i.e

$$x^6(-1024-23808x+49152x^2-5760x^3-117248x^4+189568x^5-133120x^6 \qquad (7a)$$

$$+73728x^7-92160x^8+118748x^9-73728x^{10}+16384x^{11}) = 0$$

$$y(-2560-4736y+10112y^2+12672y^3-16192y^4-1440y^5+17845y^6 -16256y^7+8320y^8-2688y^9+3072y^{10}+128y^{11}) = 0 \tag{7b}$$

In this case, equation (7a) or (7b) can be solved by obtaining the eigenvalues of the appropriate companion matrix. The solution now proceeds by solving either of the equations (7a) or (7b) and then substituting any real solutions found into (6) to find all the real common roots corresponding to each real root of (7). Alternately all the real roots of both (7a) and (7b) can be found and it can be decided by substitution which (x,y) pairs satisfy (6). The solutions corresponding to the local extrema obtained are listed in table 1a.

It should be noted that the computational effort required to find all the local extrema of this polynomial is equivalent to the effort required to solve for all of the roots of a 12th order polynomial of one variable, and is less than the computational effort required to test for its positivity by the procedure proposed by Bose and Modarressi [9].

4.2 Example 2: Global minimum of P(x,y) over H.

It is desired to find a global minimum of P(x,y) defined by (5) on some rectangular region

$$H = \{(x,y) \in \mathbb{R}^2 \,|\, x^- \le x \le x^+,\ y^- \le y \le y^+\}$$

In this case one finds all the local extrema of P(x,y) as done in table 1a and then finds the minima of P(x,y) along the boundaries. Along each of the boundaries (and along any curve in R^2) P(x,y) becomes a polynomial of a single variable and its minima can trivially be found. The results obtained for regions of different size are given in table 1b.

Since all of the minima of P(x,y) are positive it can be concluded that the polynomial P(x,y) defined by (5) is positive definite over each of the regions tested above. The global minimum of P(x,y) over H obtained in all cases is given by 4.90264×10^1

4.3 Example 3: Solution of a system of polynomial inequalities and stabilization of linear systems by output feedback.

The following example is taken from the paper by Anderson et al. [1]. Given a linear time invariant system described by

$$\begin{aligned} x &= Ax + Bu\ , \quad A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 13 & 0 \end{bmatrix} \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \\ y &= Cx \\ u &= Ky \end{aligned} \tag{8}$$

$$C = \begin{bmatrix} 0 & 5 & -1 \\ -1 & -1 & 0 \end{bmatrix}$$

TABLE 1a Local Extrema of P(x,y) (Example 1)

x	y	P(x,y)	∂P/∂x	∂P/∂y
0.00000E0	0.00000E0	5.00000E1	0.00000E0	0.00000E0
¯3.97471E¯2	¯2.49544E1	5.80405E1	1.69976E¯10	2.70006E¯13
¯5.57725E¯1	6.61279E¯1	4.92219E1	1.17684E¯14	¯4.64073E¯14
9.71122E¯1	¯4.36273E¯1	4.90264E1	¯5.04041E¯14	6.32827E¯14

TABLE 1b Local Extrema of P(x,y) along the Boundaries of Rectangular Regions (Example 2)

$$H = \{(x,y) \in \mathbb{R}^2 \,|\, x^- \le x \le x^+,\ y^- \le y \le y^+\}\ .$$

x^-,x^+,y^-,y^+	x	y	P(x,y)	∂P/∂x	∂P/∂y
$x^- = -1$	¯1.00000E0	¯1.00000E0	6.40000E1	¯2.60000E1	¯1.20000E1
	1.00000E0	¯1.00000E0	5.20000E1	6.00000E0	¯1.20000E1
$x^+ = 1$	¯1.00000E0	1.00000E0	5.20000E1	¯1.00000E1	8.00000E0
	1.00000E0	1.00000E0	5.60000E1	6.00000E0	8.00000E0
$y^- = -1$	¯1.00000E0	6.74287E¯1	5.07896E1	¯8.10237E0	¯7.66054E¯14
	1.00000E0	¯4.29445E¯1	4.90279E1	1.05223E¯1	1.85407E¯14
	3.89489E¯1	¯1.00000E0	5.02571E1	¯1.04361E¯14	¯3.51701E0
$y^+ = 1$	¯5.00000E¯1	1.00000E0	4.98125E1	2.70894E¯14	3.50000E0
$x^- = -.50$	¯5.00000E1	¯5.00000E1	1.56318E10	¯6.25775E8	¯1.24976E9
	5.00000E1	¯5.00000E1	1.56562E10	6.26245E8	¯1.25125E9
$x^+ = 50$	¯5.00000E1	5.00000E1	1.56562E10	¯6.26255E8	1.25124E9
	5.00000E1	5.00000E1	1.56308E10	6.25725E8	1.24975E9
$y^- = -50$	¯5.00000E1	4.39328E¯1	6.50129E6	¯5.15052E5	¯2.24667E¯11
	5.00000E1	4.07318E¯1	6.00153E6	4.85057E5	¯1.80058E¯11
$y^+ = 50$	¯1.99597E¯2	¯5.00000E1	5.80303E1	4.36557E¯11	4.07851E¯4
	1.99603E¯2	5.00000E1	5.79505E1	1.30967E¯10	1.17490E¯3
$x^- = -51$	¯5.10000E1	¯5.20000E1	1.90246E10	¯7.46612E8	¯1.46260E9
	5.10000E1	¯5.20000E1	1.90527E10	7.47143E8	¯1.46426E9
$x^+ = 51$	¯5.10000E1	5.20000E1	1.90527E10	¯7.47153E8	1.46425E9
	5.10000E1	5.20000E1	1.90235E10	7.46560E8	1.46259E9
$y^- = -52$	¯5.10000E1	4.39034E¯1	7.03184E6	¯5.46263E5	¯2.34303E¯11
	5.10000E1	4.07653E¯1	6.50148E6	5.15056E5	¯2.24324E¯11
$y^+ = 52$	¯1.91950E¯2	¯5.20000E1	5.80295E1	¯8.73115E¯11	3.91288E¯4
	1.91955E¯2	5.20000E1	5.79528E1	¯2.91038E¯11	1.07326E¯3
$x^- = -10^6$	¯1.00000E6	¯1.00000E6	1.00000E36	¯2.00000E30	¯4.00000E30
	1.00000E6	¯1.00000E6	1.00000E36	2.00000E30	¯4.00000E30
$x^+ = 10^6$	¯1.00000E6	1.00000E6	1.00000E36	¯2.00000E30	4.00000E30
	1.00000E6	1.00000E6	1.00000E36	2.00000E30	4.00000E30
$y^- = -10^6$	¯1.00000E6	4.23855E¯1	1.00000E24	¯4.00001E18	¯8.17099E¯3
	1.00000E6	4.23853E¯1	9.99998E23	3.99999E18	¯9.83885E¯3
$y^+ = 10^6$	¯1.00000E¯6	¯1.00000E6	5.79999E1	6.40000E2	6.98492E¯10
	1.00000E¯6	1.00000E6	5.79999E1	¯6.40000E2	¯6.98492E¯10

find the constant output feedback gains K=(v,w) which stabilize the system.

The set of polynomial inequalities arising from the Lienard-Chipart test [12] which has to be solved is:

$$v > 0 \tag{9}$$
$$w > 0$$
$$w(v-1) - (5v^2+13w) > 0$$

To find v and w which satisfy (9) by the method given in section 3, one must find a maximum of

$$P(v,w) = vw(w(v-1) - (5v^2+13w)) \tag{10}$$

over a rectangular region H, where the boundaries of H are defined by the practical limits allowed on the magnitude of the feedback gains v and w.

Local extrema of (10) are listed in table 2a. The extrema along the boundaries of H are given in tables 2b and 2c for the cases where the magnitude of the feedback gains is limited to less than or equal to 10 and 100 respectively. Since no extrema listed in tables 2a and 2b satisfy (9) there are no gains with magnitude less than 10 which stabilize the system. However in the case -100 ≤ v ≤ 100, -100 ≤ w ≤ 100 the gain K=(10.99,100) will stabilize the system. This is confirmed by noting that the eigenvalues of A+BKC are given by (-1.18±j3.12, -8.62).

TABLE 2a Local Extrema of P(v,w)

v	w	P(v,w)	∂P/∂v	∂P/∂w
0.0000E0	0.0000E0	0.0000E0	0.0000E0	0.0000E0
¯2.6000E0	0.0000E0	0.0000E0	0.0000E0	¯2.4514E¯13
¯1.4216E0	1.7294E0	¯1.0297E1	3.3884E¯13	¯6.5725E¯14
1.3716E0	3.6646E1	¯6.8457E2	¯1.0232E¯12	¯3.1264E¯13

TABLE 2b Extrema of P(v,w) along the Boundaries

	v	w	P(v,w)	∂P/∂v	∂P/∂w
	¯1.0000E1	¯1.0000E1	¯2.6000E4	1.0300E4	1.5000E3
	1.0000E1	¯1.0000E1	7.2000E4	1.9500E4	¯8.1000E3
‖v‖ ≤ 10	¯1.0000E1	1.0000E1	4.8000E4	¯1.4500E4	5.9000E3
	1.0000E1	1.0000E1	¯5.4000E4	¯1.5700E4	¯4.5000E3
‖w‖ ≤ 10	¯1.0000E1	¯1.6818E1	¯3.1114E4	1.4915E4	3.9790E¯13
	1.0000E1	3.5000E1	¯1.1025E5	¯3.8325E4	0.0000E0
	2.0384E¯1	¯1.0000E1	¯1.0404E1	¯1.0658E¯13	2.6633E0
	¯3.2705E0	¯1.0000E1	1.0381E3	¯5.6843E¯14	¯2.4348E2

TABLE 2c Extrema of P(v,w) along the Boundaries

	v	w	P(v,w)	∂P/∂v	∂P/∂w
‖v‖ ≤ 100	¯1.0000E2	¯1.0000E2	¯3.8600E8	1.2730E7	2.8500E6
	1.0000E2	¯1.0000E2	6.1200E8	1.7250E7	¯7.1100E6
‖w‖ ≤ 100	¯1.0000E2	1.0000E2	5.8800E8	¯1.6750E7	6.8900E6
	1.0000E2	1.0000E2	¯4.1400E8	¯1.3270E7	¯3.1500E6
	3.3019E¯1	¯1.0000E2	¯2.1709E3	¯8.1855E¯12	4.6221E1
	¯1.5497E1	¯1.0000E2	1.0079E6	3.6380E¯12	¯3.5644E4
	6.0641E¯1	1.0000E2	¯2.9763E3	1.8190E¯12	¯5.3631E1
	1.0994E1	1.0000E2	2.7720E5	1.3642E¯11	1.3759E4

4.4 Example 4: Minimization of a polynomial of three variables.

Find the global minimum of

$$P(x,y,z) = x^2 + (2zy-z^3)x + (z^4+z^2-2z^2y+y^2) \tag{11}$$

over a cube H in R^3 defined by

$$H = \{(x,y,z) \in \mathbb{R}^3 | -3 \le x \le 3,\ -3 \le y \le 3,\ -3 \le z \le 3\} \tag{12}$$

Solution:

The necessary conditions for the local extrema are

$$P_1(x,y,z) \triangleq \partial P/\partial x = 2x + (2zy-z^3) = 0 \tag{13}$$

$$P_2(x,y,z) \triangleq \partial P/\partial y = 2zx + (2y-2z^2) = 0$$

$$P_3(x,y,z) \triangleq \partial P/\partial z = (2y-6z^2)x + (4z^3+2z-4zy) = 0$$

These equations may be considered to be polynomials in x with coefficients in y and z. In order for these polynomials in x to have common roots the following resultants must vanish simultaneously

$$R_{12}(y,z) \triangleq (4z^2-4)y + (4z^2-4z^4) = 0 \tag{14}$$

$$R_{13}(y,z) \triangleq 4zy^2 + (8z-16z^3)y + (12z^5-8z^3-4z) = 0$$

$$R_{12}(y,z) \triangleq 4y^2 - 8z^2y + 4z^4 = 0$$

In turn, these polynomial equations may be viewed as polynomials in y with coefficients in z. They have common solutions if the folowing resultants all vanish

$$R_{(12)(13)}(z) = -64z^5 + 128z^3 - 64z = 0 \tag{15}$$

$$R_{(12)(23)}(z) = -64z^6 + 128z^4 - 64z^2 = 0$$

$$R_{(13)(23)}(z) = -1024z^8 + 2304z^6 - 1536z^4 + 256z^2 = 0$$

Equations (15) are a set of polynomial equations in one variable. They can be easily solved one at a time and the common roots obtained. The common roots are:

$$z = -1,\ 0,\ 1 \tag{16}$$

Substituting these common roots back into (14), the common values of y which satisfy all three equations (14) are obtained for each value of z. The solutions of (14) are:

$$(y,z) = (0,0),\ (0,-1),\ (0,1),\ (2,1),\ (2,-1) \tag{17}$$

Local extrema finally obtained by substituting (17) into (13) are listed in table 3a.

To find the global minimum over H, minimizations must be performed over the "faces" and "edges" of the cube and the function must be evaluated at its eight corners. Results of these minimizations (carried out as in example 1) are summarized in table 3b. From table 3b it can be seen that the global minimum over H is -63 and occurs at (x,y,z) = (3,0,3) and at (x,y,z) = (-3,0,-3).

TABLE 3a Local Extrema of P(x,y,z) (Example 4)

	x	y	z	P(x,y,z)	∂P/∂x	∂P/∂y	∂P/∂z
local extrema	0	0	0	0	0	0	0
	1	0	1	1	0	0	0
	-1	0	1	1	0	0	0
	-1	2	1	1	0	0	0
	1	2	-1	1	0	0	0

TABLE 3b Local Extrema of P(x,y,z) along the Boundaries -3 ≤ x ≤ 3, -3 ≤ y ≤ 3, -3 ≤ z ≤ 3

	x	y	z	P(x,y,z)	∂P/∂x	∂P/∂y	∂P/∂z
"corners"	¯3.00000E0	¯3.00000E0	¯3.00000E0	¯5.40000E1	6.60000E1	¯6.00000E0	3.00000E1
	3.00000E0	¯3.00000E0	¯3.00000E0	3.78000E2	7.80000E1	¯4.20000E1	¯3.30000E2
	¯3.00000E0	3.00000E0	¯3.00000E0	¯5.40000E1	3.00000E1	6.00000E0	6.60000E1
	3.00000E0	3.00000E0	¯3.00000E0	1.62000E2	4.20000E1	¯3.00000E1	¯2.22000E2
	¯3.00000E0	¯3.00000E0	3.00000E0	3.78000E2	¯7.80000E1	¯4.20000E1	3.30000E2
	3.00000E0	¯3.00000E0	3.00000E0	¯5.40000E1	¯6.60000E1	¯6.00000E0	¯3.00000E1
	¯3.00000E0	3.00000E0	3.00000E0	1.62000E2	¯4.20000E1	¯3.00000E1	2.22000E2
	3.00000E0	3.00000E0	3.00000E0	¯5.40000E1	¯3.00000E1	6.00000E0	¯6.60000E1
"edges"	¯3.00000E0	0.00000E0	¯3.00000E0	¯6.30000E1	4.80000E1	0.00000E0	4.80000E1
	3.00000E0	0.00000E0	3.00000E0	¯6.30000E1	¯4.80000E1	0.00000E0	¯4.80000E1
	¯3.00000E0	3.00000E0	8.16890E¯1	2.65421E1	¯9.81110E1	9.56672E0	¯1.47882E¯13
	¯3.00000E0	3.00000E0	1.14174E0	1.56015E0	¯2.12623E0	¯3.45753E0	¯3.27738E¯13
	3.00000E0	3.00000E0	8.16890E¯1	2.65421E1	9.81110E0	9.56672E0	1.47882E¯13
	3.00000E0	3.00000E0	¯1.14174E0	1.56015E0	2.12623E0	¯3.45753E0	3.27738E¯13
"faces"	¯3.00000E0	0.00000E0	0.00000E0	9.00000E0	¯6.00000E0	0.00000E0	0.00000E0
	3.00000E0	0.00000E0	0.00000E0	9.00000E0	6.00000E0	0.00000E0	0.00000E0
	0.00000E0	¯3.00000E0	0.00000E0	9.00000E0	0.00000E0	¯6.00000E0	0.00000E0
	0.00000E0	3.00000E0	0.00000E0	9.00000E0	0.00000E0	6.00000E0	0.00000E0
	3.85968E¯1	3.00000E0	1.79311E0	3.11261E0	9.41469E¯14	9.53667E¯1	¯1.42109E¯13
	¯3.85968E¯1	3.00000E0	¯1.79311E0	3.11261E0	¯9.76996E¯14	9.53667E¯1	1.49214E¯13
	¯1.86565E0	3.00000E0	1.20475E0	3.68870E¯1	8.88178E¯15	¯1.39811E0	¯3.50830E¯14
	1.86565E0	3.00000E0	¯1.20475E0	3.68870E¯1	¯8.88178E¯15	¯1.39811E0	4.13003E¯14

5. SUMMARY AND CONCLUSIONS.

A procedure which finds the global minimum and all of the local extrema of a multivariable polynomial over a closed hypercube in R^n has been presented. It has been demonstrated how this procedure may be used to test for positivity of polynomials over closed regions in R^n and to solve sets of multivariable polynomial inequalities. The proposed procedure has been tested on two examples taken from Bose and Modarressi [9] and Anderson and Scott [2] and in both cases the computational effort required is less than that required in [9],[2].

A large class of practical engineering problems which may be formulated as a set of polynomial inequalities over a closed subset of R^n lends itself to the solution by the procedure presented in this paper. The procedure is especially useful if there are practical limitations on the magnitude which certain variables may take.

At present, the practical usefulness of the above algorithm is limited by the lack of efficient computer algorithms for the solution of systems of polynomial equations. Using the simple APL programs written by the authors for symbolic polynomial manipulation, the computational effort becomes excessive for four or more variables.

ACKNOWLEDGEMENT.

The authors would like to thank Mr. George Klein for long and fruitful discussions and many helpful suggestions.

REFERENCES.

[1] B. Anderson, N. Bose and E. Jury, "Output feedback stabilization and related problems - solution via decision methods", IEEE Trans. Auto. Control, Vol.AC-20, No. 1, Feb. 1975, pp. 53-66.

[2] B. Anderson and R. Scott, "Output feedback stabilization - solution by algebraic geometry methods", The University of Newcastle Technical Report EE7530, August 1975.

[3] A.Taylor, Advanced Calculus, Blaisdell Publishing Co., Waltham Mass., 1955.

[4] J.Moses, "Solution of systems of polynomial equations by elimination", Comm. of the ACM, Vol. 9, No. 8, August 1966.

[5] G.E. Collins, "Computer algebra of polynomials and rational functions", Amer. Math. Monthly, Vol. 80, Sept 1973, pp.725-755.

[6] J.L. Coolidge, A Treatise on Algebraic Plain Curves, Oxford 1931.

[7] H.H. Rosenbrock, "An automatic method for finding the greatest or least value of a function", Computer J., Vol.3, No. 3, October 1960, pp. 175-184.

[8] E.J. Davison and P. Wong, "A robust conjugate-gradient algorithm which minimises L-functions", Automatica, Vol.11, 1975, pp.297-308.

[9] N.K. Bose and A.R. Modarressi, "General procedure for multivariable polynomial positivity test with control applications", To be published.

[10] N. Jacobson, Lectures in Abstract Algebra, Vol. III, Van Nostrand, Princeton, N.J., 1964.

[11] B. Smith et al., Matrix Eigensystem Routines - Eispack Guide, Springer Verlag 1970.

[12] F.R. Gantmacher, Applications of the Theory of Matrices, Interscience, New York 1959.

[13] W.S. Gesing and E.J. Davison, "Global minimization of polynomials of n variables with control applications", Dept. of Electrical Engineering, University of Toronto, System Control Report No. 7616 Sept. 1976.

ON THE BEHAVIOR OF OPTIMAL SAMPLED DATA REGULATORS DRIVEN BY STOCHASTIC INPUT

DINESH A KESKAR
Graduate Research Assistant

GARY L. SLATER
Associate Professor
Department of Aerospace Engineering & Applied Mechanics
University of Cincinnati
Cincinnati, OH 45221

This paper examines the stochastic response of continuous systems which are sampled and controlled continuously. The approach taken is to discretize the system and performance index so as to obtain the optimal feedback gains from the optimal discrete regulator theory. The effect of sampling on the state error and control covariance matrices is determined. Analytical and numerical results are obtained for several low order systems in order to demonstrate the results. For simple first and second order systems the stochastic response is similar to the behavior of the optimal cost function as determined by Levis, et-al. [1]. Thus for example the state covariance of a stable first order system approaches a constant value with increasing sample rate while the control covariance monotonically decreases. For the unstable first order system the covariance show a divergent behavior. Similarly analogous to the cited results the state covariance of the harmonic oscillator is unbounded when the sampling rate is a multiple of half the sampling period. The control covariance in this case is unbounded only at even multiples of the Nyquist frequency with a relative minimum at the odd multiples.

The methods are applied to the determination of sample rate for the longitudinal control of a small flexible aircraft. The primary goal of the controller is to minimize the acceleration loads felt by the pilot. Several approaches to the determination of the required sample rate are investigated. The classical approach is to determine a set of feedback laws completely ignoring the structural motion. If the effects of the structural motion are significant a digital notch filter is inserted into the control loop to filter commanded control deflections to the aircraft. These conventional approaches are compared with the optimal filtering solutions which include the structural motion in their model. For all cases the effect of sampling rate on the mean square response is obtained to determine the sampling rate requirement.

Even though the particular aircraft model considered here is reasonable rigid the results indicate a significant improvement in system response (a factor of two in RMS response) is obtained by including the structural motion in the control model. The addition of the notch filter yields only a slight improvement in response at the faster sampling rates and in fact degrades the response at low sampling rates.

[1] Levis, et.al. "On the Behavior of Optimal Linear Sampled-Data Regulators" Int. J. Control, 1971, Vol. 13, No. 2

OPTIMAL SAMPLING FOR DATA ACQUISITION SYSTEMS

RONALD L. VAN WIEREN
Lear Siegler, Inc.
Grand Rapids, Michigan

ROBERT A. SCHLUETER
Department of Electrical
and Systems Science
Michigan State University
East Lansing, Michigan

ABSTRACT

The optimal sampling problem is formulated for data acquisition systems used in control. A performance index is used that measures both the error due sampling in a digital data acquisition system for a particular sampling criterion and the costs of implementing that criterion. A computational algorithm for computing the optimal sampling times sequence for periodic, adaptive, and aperiodic criteria is developed and used on an example problem. The performance of periodic, adaptive, and aperiodic criteria are compared for different inputs and performance indices on this example.

I. INTRODUCTION

The design of sampling criteria has always been based on control performance, signal sampling efficiency, cost of implementation, and ease of design and analysis. Periodic sampling criteria have often been used because periodic sampled-data single-output control systems can then be easily designed using transform techniques. For periodic sampling, the sampling rate is generally constrained to satisfy the Nyquist sampling theorem. As a result, the sampling rate is selected to be greater than twice the system bandwidth in order to minimize aliasing and thus provide reasonable control performance. The actual choice of a sampling rate is chosen by a tradeoff between control performance and the economic cost for instrumentation and communications hardware. The continuous-time controller is implemented if the cost for communicating data is low and computer processing is not required to implement the system design.

Modern digital data acquisition systems often are implemented with microprocessors for signal processing, coding, and multiplexing. The same microprocessors could be used to adjust the triggering rate of the sample-and-hold mechanism if the improvement in performance could be justified against the additional cost.

Much effort has been expended in developing nonperiodic sampling criteria that could improve performance at reasonable cost. Multirate and asynchronous sampling criteria [1] have been proposed for multi-input multi-output systems where different signals can have widely varying characteristics. Adaptive sampling criteria [2-9] have also been developed which vary the sampling rate with respect to changes in the monitored signals. If the signals used for sampling rate adjustment accurately reflect system dynamics and their changes due to control inputs, disturbances, and plant parameter variation, the sampling criterion could be effectively tuned to the system. Thus, one sampling rule could be used to sample all the output signals monitored by a data acquisition system, thus permitting improved performance at a minimum additional cost.

The function of a sample-and-hold mechanism in a data acquisition system is to sample often enough and at appropriate sampling times so that the error between the continuous signal being observed and the piecewise constant approximation, which is the output of the sample-and-hold mechanism, is as small as possible. A performance index for optimal sampling in data acquisition systems should not only measure this error in a sample-and-hold approximation for any sampling criterion, but also the cost of implementing that criterion. This cost of implementation should reflect the costs for designing or tuning the sampling criterion, the costs of communicating the sampled data and any hardware costs for instrumentation, computer, or communication equipment associated with implementing that sampling criterion.

Hsia [10] has recently shown that almost all adaptive sampling criteria known [2-9] can be derived analytically by minimizing a Taylor series approximation of the performance index

$$\underline{J}_i = \frac{1}{T_i^{k_1}} \int_{t_i}^{t_i+T_i} |s(t) - s(t_i)|^{k_2}\, dt + \omega_1 e^{-\alpha T_i} \tag{1}$$

with respect to the sampling intervals T_i for $i = 0,1,...,N-1$.

The first term is a weighted integral of the error between the continuous signal $s(t)$ and the last sampled measurement $s(t_i)$ over the entire interval $[t_i, t_i+T_i]$. This measure penalizes both size and duration of this error. The second term was called "cost of sampling" and was not intended to actually measure costs of implementation, but rather to avoid trivial solutions (T_i=0). The parameters (ω,α,k_1,k_2) can take on the following values:

$$\omega_1,\ \alpha\ \varepsilon(0,\infty)$$

$$k_1\ \varepsilon\{-1,0,1\};\ k_2\ \varepsilon\{1,2\}.$$

The signal $s(t)$ is approximated by the truncated Taylor series

$$s(t) \cong s(t) + \dot{s}(t_i)(t-t_i) \qquad t \geq t_i \tag{2}$$

and the "cost for sampling" is approximated by either

$$\omega_1 e^{-\alpha T_i} \approx \begin{cases} \omega_1[1 - \alpha T_i] \\ \omega_1[1 - \alpha T_i + (\alpha T_i)^2/2] \end{cases} \tag{3}$$

before the actual sampling rule is selected. The sampling rule for a particular set of parameters $(\omega_1,\alpha,k_1,k_2)$ can be determined by minimizing (1) with respect to T_i having substituted (2) and (3). The rules developed by this procedure express the sampling interval length T_i as a function of performance index parameters $(\omega_1, \alpha, k_1, k_2)$ and the signal and its derivatives at the sampling time t_i.

These adaptive sampling rules select sampling times sequentially and base the selection on an estimate (2) of the change in $s(t)$ using measurements $s(t_i)$ and $\dot{s}(t_i)$. The performanc of these rules depends not only on the accuracy of the approximati (2) and (3), but also on the validity of the performance index used.

Smith [11] made a performance comparison of periodic and adaptive criteria that indicated any adaptive sampling rules specified by k_1 and k_2 must be tuned by adjusting parameters α and ω_1 for the particular system, control inputs, etc., if the adaptive criterion were to outperform periodic sampling. A conclusion that can be made from this study is that sampling rules should be designed based on the system, control inputs and disturbances in the system, and initial conditions at the time sampling is initiated rather than tuning a criterion that has been designed without this information. A second conclusion from this study is that a design stage is necessary if an efficient, effective sampling criterion is to be developed. Smith's study also indicated that adaptive rules would generally sample at extremes if constraints on the minimum and maximum sampling interval were imposed to insure good performance. This result could occur if the signal approximation (2) were not accurately predicting changes in $s(t)$. Another disadvantage of these adaptive rules is that sampling rate adjustment must be based on only one signal, where the resulting sampling criterion could be used in sampling several signals in a system. This limitation of these adaptive sampling criteria [10] was caused by using a performance measure (1) that considered the sampling error in only one signal. Finally, the cost of implementation used was selected based on mathematical convenience, as in the derivation of these adaptive sampling criteria, but should be based on the actual design, hardware, and communication costs for implementing a particular sampling criterion.

Optimal sampling problems for data acquisition in control systems are formulated in an effort to develop sampling criteria that are: (1) based on apriori information about the system and are not based on some noise-sensitive estimate of changes in the signal; (2) based on signal representation quality in all signals of interest and not just a single signal; and (3) based on an accurate measure of the costs for implementing sampling criteria. The optimal aperiodic sampling times sequence will be selected together in a block rather than sequentially, thus permitting selection of an optimal number of sampling times also.

The problem is formulated in Section II, where a model of the control system and data acquisition system is stated and the development of a performance index is undertaken. The algorithm for computing the sampling times sequence for optimal aperiodic sampling and the optimal number of samples for periodic and adaptive sampling criteria is presented in Section III. This algorithm is used to evaluate the performance of these criteria for a system with different inputs and different performance indices in Section IV. Conclusions are presented in Section V.

II. PROBLEM FORMULATION

Consider the linear time-invariant system

$$\dot{\underline{x}}(t) = \underline{A}\ \underline{x}(t) + \underline{B}\ \underline{u}(t) + \underline{\omega}(t) \tag{4}$$

with outputs

$$\underline{y}(t) = \underline{C}\ \underline{x}(t) \tag{5}$$

and sampled-data control

$$\underline{u}(t) = \underline{u}_i = -\underline{G}\ \underline{y}(t_i) + \underline{z}(t_i) \qquad t\varepsilon[t_i, t_{i+1}] \tag{6}$$

for i=0,1,...,N-1 where $\underline{z}(t)$ is the known input, $\underline{\omega}(t)$ is the noise disturbance with

$$E\{\underline{\omega}(t)\} = \underline{\omega}_o \tag{7}$$

$$\mathrm{cov}\{\underline{\omega}(t), \underline{\omega}(t)\} = \underline{\psi}(t)\delta(t-\tau) \tag{8}$$

and $\{t_i\}_{i=0}^{N-1}$ are the sampling times at which measurements of $\underline{y}(t)$ and $\underline{z}(t)$ are made and then used to construct control (6). The initial time t_o and $t_f = t_N$ is known and the initial state $\underline{x}(t_o)$ is random with

$$E\ \{\underline{x}(t_o)\} = \underline{m}(t_o) \qquad \mathrm{cov}\{\underline{x}(t_o), \underline{x}(t_o)\} = \underline{V}(t_o) \tag{9}$$

where $\underline{x}(t_o)$ and $\underline{\omega}(t)$ for all $t\varepsilon[t_o, t_f]$ are independent.

The data acquisition system will sample $\underline{y}(t)$ and $\underline{z}(t)$ at t_i and then transmit $\underline{y}(t_i)$, $\underline{z}(t_i)$, and t_{i+1} to the controller so that control (6) can be implemented over $[t_i, t_{i+1})$ for i=0,1,..., N-1. The sampling times sequence $(\{t_i\}_{i=1}^{N-1}, N)$ used by the data acquisition system will be selected off-line based on minimizing the sampling error on a signal vector

$$\underline{s}(t) = \underline{E}\ \underline{y}(t) + \underline{F}\ \underline{z}(t) \tag{10}$$

where the components of this vector can include any element of the output $\underline{y}(t)$, input $\underline{z}(t)$, or control $-\underline{G}\ \underline{y}(t) + \underline{z}(t)$ vectors. The matrices $\underline{E}$ and $\underline{F}$ would be selected so that the appropriate elements of these signal vectors would be used in the design of the aperiodic criterion. The sample-and-hold approximation of $\underline{s}(t)$ is defined as

$$\hat{\underline{s}}(t) = \underline{s}(t_i) \qquad t\varepsilon[t_i, t_{i+1}) \quad i = 0,1,\ldots,N-1 \tag{11}$$

where the sampling times $\{t_i\}_{i=1}^{N-1}$ must satisfy

$$t_0 < t_1 < t_2 \ldots < t_{N-1} < t_N = t_f \tag{12}$$

The sampling interval and sampling interval sequence are defined as $T_i = t_{i+1} - t_i$ and $(N,\underline{T})$, respectively, where $\underline{T}^T = (T_0,T_1,\ldots,T_{N-1})$.

An optimal periodic sampling can be specified by selecting the number of sample intervals, N, to satisfy

$$N_{min} \leq N \leq N_{max} \tag{13}$$

and imposing the equality constraint knowing t_f and t_o

$$\underline{g}(\underline{T}) = \frac{t_f - t_o}{N} - T_i = 0, \ i = 0,1,\ldots,N-1 \tag{14}$$

where N_{max} is determined by hardware cost specifications and N_{min} by signal representation quality. Optimal periodic sampling is obtained through the minimization of a data acquisition system performance index with respect to N satisfying (13) and (14), thus yielding the optimal number of samples N_p^*. Periodic sampling is specified by a fixed N such that

$$N_{min} = N = N_{max} \tag{15}$$

and requiring (14) be satisfied for the N chosen.

An optimal aperiodic sampling criterion can be specified by sampling interval constraints of the form

$$0 < T_{i_{min}} \leq T_i \leq T_{i_{max}} \tag{16}$$

and given t_f and t_o

$$\underline{g}(\underline{T}) = \sum_{i=0}^{N-1} T_i - [t_f - t_o] = 0 \tag{17}$$

where $T_{i_{min}}$ is determined by the data acquisition system cost specifications and $T_{i_{max}}$ by system stability or worst case signal representation quality. The number of sample intervals, N, is chosen to satisfy (13). Optimal aperiodic sampling is obtained through minimization of a data acquisition system performance index with respect to both N and $\underline{T}$ satisfying (13), (16), and (17) to yield the optimal number of sampling intervals, N_o^*, and the optimal sampling interval sequence, $\underline{T}^*$. Suboptimal aperiodic sampling specified a fixed N satisfying (15). Suboptimal aperiodic sampling criterion is obtained through minimization of a data acquisition system performance index with respect to $\underline{T}$, with N satisfying (15), (16), and (17). Finally, aperiodic sampling is specified when N and $\underline{T}$ are chosen to satisfy (15), (16), and (17) with no optimization on N or $\underline{T}$.

Five sampling criteria have been specified. Each of the optimal criteria require the minimization of a data acquisition system performance index (S) with respect to certain variables. The form of the performance index (S) used will be similar to (1) in that both signal representation quality (J) and cost of implementation (C) will be measured so that

$$S = J + C. \tag{18}$$

The signal representation performance index should measure the error between signal $\underline{s}(t)$ and the piecewise constant approximation $\hat{\underline{s}}(t)$ over the entire observation interval $[t_o, t_f]$ rather than just one sampling interval $[t_i, t_{i+1})$ because the sampling interval sequence will be selected together and not sequentially. Moreover, the performance of any criterion makes more sense over some observation interval rather than a sampling interval to-sampling interval basis. The index should then measure the average signal representation quality for all sample functions of the process $\underline{s}(t)$ because the optimal sampling interval sequence must be computed off-line and thus should not be based on a single sample function. The performance index (1) should thus be restricted to be quadratic (k_2=2) for mathematical convenience. Finally, the signal representation performance index should measure signal representation quality for all elements of the input, output, and control error signal vectors of interest rather than just one signal chosen to represent the dynamics in the system. The signal representation performance index chosen has the form:

$$J = E\{ \sum_{i=0}^{N-1} \frac{1}{T_i^{k_i}} \int_{t_i}^{t_{i+1}} || \underline{s}(t) - \underline{s}(t_i) ||_{\underline{S}}^2 \, dt \}. \tag{19}$$

The matrix $\underline{S}$ is assumed to be a positive semidefinite q x q symmetric matrix, T_i is assumed to satisfy the appropriate sampling interval constraints, and E{•} is the expectation operator taken over the random variables found in the system which generates the continuous time signal vector $\underline{s}(t)$. The functional J penalizes sampling errors with respect to the mean of the integral if k_1=1, the time-weighted integral if k_1=-1, or the unweighted integral if k_1=0.

The cost of implementation (C) in this data acquisition system performance index should measure the cost (C_1) of the additional instrumentation, communication, or computer hardware required to implement a criterion; the cost (C_2) for designing or tuning a sampling criterion in order to achieve its best possible performance; and the cost (C_3) for communication of the sampled measurements $\underline{z}(t_i)$, $\underline{y}(t_i)$, and t_{i+1} for i=0,1,...,N-1 to the controller for implementation of control (6). Thus the cost of implementation has the form

$$C = C_1 + C_2 + C_3. \tag{20}$$

The cost for additional hardware (C_1) can be neglected if the sampling criterion is to be implemented on an existing or specified set of hardware. However, if the instrumentation, computer, and communication hardware to be implemented as part of the data acquisition system will depend on the sampling criterion selected, this cost term must be included. An accurate determination of the hardware cost could be developed for a particular application and any sampling criterion considdered for implementation. Although the functional form of this cost term may vary for different applications, the hardware cost would generally be monotonically decreasing function of the minimum sampling period $\min_i\{T_i\}$ and a monotonically increasing function of the on-line memory and computational requirement for a particular criterion. If on-line computation and memory is not required, a possible functional form for this hardware cost term is

$$C_1 = \omega_1 \, \varepsilon^{-\alpha \min_i \{T_i\}} \tag{21}$$

where ω_1 is a constant proportional to the cost for implementing an analog data acquisition system and $(-\alpha)$ is a constant that reflects the marginal cost for increasing the minimum sampling period.

The design cost term (C_2) should not only measure the off-line computational costs for selecting the sampling times sequence $\{t_i\}_{i=0}^{N-1}$ for a criterion having N sampling times, but also the computational costs for evaluating the system performance index using that sequence. This measure apportions the total design costs for computing sampling times sequence $\{t_i\}_{i=0}^{N-1}$ and system performance index values S for all N satisfying (13) so that the costs for any N are just equal to the computational costs for that N. Thus, the design cost C_2 can be expected to increase with N since the number of variables to be selected increases with N.

Computational costs are generally proportional to the CPU time (U) required for any task and the memory (M) required by the program performing the task. A common measure of computer utilization used for determining charges for computer usage is the product $(M \cdot U)$ measured in bit-seconds. Thus, the design cost for a particular criterion using N samples is

$$C_2(N) = \omega_2 \, M \cdot U \tag{22}$$

where ω_2 is the cost per bit-second on the computer system used for the off-line design effort. The CPU time U should reflect the time required for computing a sequence $\{t_i\}_{i=0}^{N-1}$ for a particular criterion with N sampling times and the time required to evaluate S for that sequence. The memory (M) should be the memory requirements of the program used to perform the above tasks.

The communication system utilization cost (C_3) should only be included if the communication link is time-shared with other users. If the link is dedicated to the data acquisition system, the cost for communication should be totally accounted for in hardware cost (C_1) and the utilization cost C_3 can be omitted. The communication system utilization cost for transmitting the data $\underline{y}(t_i)$, $\underline{z}(t_i)$, and in some cases t_{i+1}, for $i=0,1,\ldots,N-1$ (required to construct the control (6)) should depend on word size (W) for each piece of data, the total number of words K(N) required to transmit this data, and the communication capacity R of the time-shared communication link. The "connect time" $\left(\frac{K \cdot W}{R}\right)$ is a common measure of the utilization of a communication link by any one user and is often used as a basis for determining the charge for such use. Thus, the communication system utilization cost term has the form

$$C_3 = \omega_3 \left(\frac{K \cdot W}{R}\right) \tag{23}$$

where ω_3 is proportional to the cost per second on the communication link used in the particular application. The number of bits/word (W) depends on the accuracy of the data required and the coding performed. The number of words required is

$$K = \begin{cases} (m + r)N & \text{for periodic} \\ (m + r + 1)N & \text{for aperiodic and adaptive} \end{cases}$$

because only the measurement $\{\underline{z}(t_i)\}_{i=0}^{N-1}$ and $\{\underline{y}(t_i)\}_{i=0}^{N-1}$ must be transmitted for periodic sampling, but these measurements and the sampling times $\{t_i\}_{i=1}^{N}$ must be transmitted for adaptive and aperiodic sampling. The channel capacity R must be large enough so that $(m + r + 1)W$ bits, which represent $\underline{z}(t_i)$, $y(t_i)$, and t_{i+1}, can be transmitted in $\min\limits_i\{T_i\}$ seconds.

The optimal aperiodic sampling for the data acquisition problem can now be stated formally.

Given system (4) with output (5) and control (6), where $\underline{z}(t)$ is a known programmed control, and $\underline{\omega}(t)$ and $\underline{x}(t_o)$ are the disturbance and initial state respectively with statistics (7,8,9), determine the sampling interval sequence $(N,\underline{T})$ that minimizes

$$\begin{aligned} S &= J + C_1 + C_2 + C_3 \\ &= E\{\sum_{i=0}^{N-1} \frac{1}{k_i T_i} \int_{t_i}^{t_i+T_i} ||s(t) - \underline{s}(t_i)||^2_{\underline{S}}\, dt\} \\ &\quad + \omega_1 \varepsilon^{-\alpha \min\limits_i\{T_i\}} + \omega_2\, M \cdot U + \omega_3 \frac{W \cdot K}{R} \end{aligned} \tag{24}$$

subject to the sampling constraints for periodic (13,14) and aperiodic (13,16,17) criteria.

III. COMPUTATIONAL ALGORITHM

The signal representation performance index (J) and data acquisition performance index (S) must now be expressed as functions of the sampling interval sequence $(\underline{T},N)$.

The signal representation performance index can be expressed as

$$J = \sum_{i=0}^{N-1} \frac{1}{k_i T_i} E\{\int_{t_i}^{t_i+T_i} [\underline{E}\ \underline{C}\ \hat{\underline{x}}(t) + \underline{F}\ \hat{\underline{z}}(t)]^T\ \underline{S}\ [\underline{E}\ \underline{C}\ \hat{\underline{x}}(t) + \underline{F}\ \hat{\underline{z}}(t)]dt \tag{25}$$

where

$$\left.\begin{aligned} \hat{\underline{x}}(t) &= \underline{x}(t) - \underline{x}(t_i) \\ \hat{\underline{z}}(t) &= \underline{z}(t) - \underline{z}(t_i) \end{aligned}\right\} \quad t\ \varepsilon [t_i, t_{i+1});\ i = 0,1,\ldots,N-1.$$

The performance index $J = J(\underline{T},N)$ becomes

$$J(\underline{T},N) = \sum_{i=0}^{N-1} \frac{1}{T_i^{k_i}} \int_{t_i}^{t_i+T_i} \Big([\underline{E}\ \underline{C}\ \hat{\underline{m}}(t) + \underline{F}\ \hat{\underline{z}}(t)]^T\ \underline{S}\ [\underline{E}\ \underline{C}\ \hat{\underline{m}}(t) + \underline{F}\ \hat{\underline{z}}(t)] + \mathrm{Tr}\{\underline{C}^T\ \underline{E}^T\ \underline{S}\ \underline{E}\ \underline{C}\ \hat{\underline{V}}(t)\} \Big) dt \qquad (26)$$

where $\mathrm{Tr}\{\cdot\}$ is the trace operator and the mean and variance of $\underline{x}(t)$ are

$$\underline{m}(t) = [\underline{\Phi}(t,t_i) - \underline{D}(t,t_i)\underline{B}\ \underline{G}\ \underline{C}]\ \underline{m}(t_i) + \underline{D}(t,t_i)[\underline{B}\ \underline{z}(t_i) + \underline{\omega}_o]$$

$$\underline{V}(t) = [\underline{\Phi}(t,t_i) - \underline{D}(t,t_i)\underline{B}\ \underline{G}\ \underline{C}]\underline{V}(t_i)[\underline{\Phi}(t,t_i) - \underline{D}(t,t_i)\underline{B}\ \underline{G}\ \underline{C}]^T + \int_{t_i}^{t} \underline{\Phi}(t,\tau)\ \underline{\psi}(\tau)\ \underline{\Phi}^T(t,\tau)d\tau \qquad (27)$$

for $t \varepsilon [t_i,t_{i+1})$ and $i = 0,1,\ldots,N-1$. The data acquisition performance (S) can then be expressed as a function of the sampling interval sequence $(\underline{T},N)$.

The performance index for periodic sampling can be evaluated for any N and

$$T_i = \frac{t_f - t_o}{N} = T \qquad i = 0,1,\ldots,N-1 \qquad (30)$$

to obtain a performance $S_p(N)$. The optimal periodic sampling criterion is determined by determining the value N that minimizes $S_p(N)$ subject to (13). Since N will be small and since the shape of $S_p(N)$ is of interest, this minimization will be performed graphically by plotting $S_p(N)$. The minimization could be performed using an integer programming algorithm.

The optimal aperiodic and suboptimal aperiodic sampling criteria will be evaluated for the case where

$$T_{i_{min}} = -\infty \qquad T_{i_{max}} = \infty \qquad i = 0,1,\ldots,N-1$$

and thus the inequality constraint (16) can be ignored. The suboptimal aperiodic sampling criterion will be computed by minimizing S with respect to $\underline{T}$ subject to (17) for a specified number of samples N to obtain $S_o(N) = S(\underline{T}^*,N)$. This minimization of S subject to (17) can be performed by substituting

$$T_{N-1} = t_f - t_o - \sum_{j=0}^{N-2} T_j \qquad (31)$$

and then minimizing S with respect to $T_0,T_1,\ldots,T_{N-2}$ using a Powell [12] algorithm.

The optimal aperiodic sampling criterion $(N_0^*,\underline{T}^*)$ can then be determined by minimizing $S_o(N)$ with respect to N satisfying (13). This minimization will again be carried out graphically after $S_o(N)$ is plotted versus N because N will be small and the shape of $S_o(N)$ is of interest. An integer programming algorithm could be used to determine $(N_o^*, \underline{T}^*)$ and $S_o(N_o^*)$ if the algorithm calls for a subroutine that evaluates $S_o(N)$ through minimization of S with respect to $\underline{T}$.

V. REFERENCES

1. R.E. Kalman and J.E. Bertram, "A Unified Approach to the Theory of Sampling Systems," Proceedings of the IRE, Vol. 46, 1958.

2. R.C. Dorf, M.C. Farren, and C.A. Phillips, "Adaptive Sampling for Sampled-Data Control Systems," IRE Trans. on Automatic Control, Vol. AC-7, pp. 38-47, 1962.

3. R. Tomovic and G.A. Bekey, "Adaptive Sampling Based on Amplitude Sensitivity," IEEE Trans. on Automatic Control, Vol. AC-11, pp. 282-284, April 1966.

4. D. Ciscator and L. Mariani, "On Increasing Sampling Efficiency by Adaptive Sampling," IEEE Trans. on Automatic Control, Vol. AC-12, p. 318, June 1967.

5. A.W. Bennett and A.P. Sage, "Discrete System Sensitivity and Variable Increment Sampling," Proceedings Joint Automatic Control Conference, pp. 603-612, June 1967.

6. K.E. Tait, "Evaluation of Signal Dependent Sampling," Intl. Journal on Control, Vol. 4, No. 4, pp. 201-239, 1966.

7. G.A. Bekey and R. Tomovic, "Sensitivity of Discrete Systems to Variation of Sampling Intervals," IEEE Trans. on Automatic Control, (short paper), Vol. AC-11, pp.284-287, 1966.

8. J.R. Mitchell and W.L. McDaniel, Jr., "Adaptive Sampling Technique," IEEE Trans. on Automatic Control, Vol. AC-14, pp. 200-201, April 1969.

9. S.C. Gupta, "Increasing the Sampling Efficiency for a Control System," IEEE Trans. on Automatic Control, (correspondence), Vol. AC-8, pp. 263-264, July 1963.

10. T.C. Hsia, "Analytic Design of Adaptive Sampling Control Law in Sampled-Data Systems," IEEE Trans. on Automatic Control, Vol. AC-19, pp. 39-42, February 1974.

11. M.L. Smith, Jr., "An Evaluation of Adaptive Sampling," IEEE Trans. on Automatic Control, Vol. AC-16, pp. 282-284, June 1971.

12. M.J.D. Powell, "An Efficient Method for Finding the Minimization of a Function of Several Variables Without Calculating Derivatives," Computer Journal, Vol. 7, pp. 155-162, 1964.

SOLUTION OF MULTISENSOR TRACKING PROBLEMS ON ASSOCIATIVE PROCESSORS

DR. CHARLES L. MOREFIELD
ORINCON Corporation, La Jolla, California 92037

ABSTRACT

Multisensor, multitarget tracking can be accomplished through the use of decision-directed estimation schemes. This approach has generally been avoided in the literature due to the tremendous computational load it imposes. In this paper we demonstrate that a combination of associative processing hardware and 0-1 integer programming software can reduce the computational burden of such decision-directed estimation problems to a remarkably low level.

1. INTRODUCTION

Multisensor tracking problems arise in a variety of situations, including sea surveillance, ballistic missile defense, and air traffic control, and involve a variety of sensors, including radar, sonar, and infrared. The distinguishing characteristic of these problems is the difficulty encountered in assigning the available measurements to individual targets for tracking purposes. Stated in the simplest possible terms, multisensor tracking involves a set of measurements from several sensors, a target motion model and a decision process. The measurements might be, for example, azimuth and range data from several radars scanning the aircraft in a busy air corridor. Target motion models are often expressed in terms of predictions from a Kalman filter. Individual tracks are formed through a decision process which matches filter predictions to the available measurements in some manner. Given a large number of measurements from various sensors, we must determine which measurements are associated with the same target, which measurements are false alarms, how many targets are present, and what the position, course, and speed is of each target.

Data produced by the sensors may be collected in tabular form as shown below:

Measurement #	Time	Measurement	Sensor #	Target #
1	t_1	z^1	s_1	?
2	t_2	z^2	s_2	?
3	t_3	z^3	s_3	?
4	t_4	z^4	s_4	?
5	t_5	z^5	s_5	?
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
N	t_N	z^N	s_N	?

TABLE 1. Data Collected by Multiple Sensors.

In the above table, N measurements $\{z^1, z^2, \ldots z^N\}$ are taken at a number of points in time t_i by various sensors s_i. Some of the time points may be the same. Also, since one sensor may produce many measurements, many of the sensor numbers s_i may be the same.

As the table indicates, the key information lacking in a multisensor, multitarget environment is a label identifying the target associated with each measurement. For this discussion, the target identity of each measurement is simply a number that identifies which measurements should be associated with a particular target. Since some measurements may be due to random noise, it is also desirable to find and label such measurements as false alarms.

Data association techniques that are useful in situations such as this can be found in both the pattern recognition and multitarget tracking literature [1-15] . In the pattern recognition literature [15], the problem is often referred to as an "unsupervised" pattern recognition problem. Using statistical models for the data*, the data is partitioned into the clusters that best fit these models. The "best fit" is typically determined through the use of Bayesian decision theory.

A number of investigators have developed approaches to multitarget tracking that are useful in these circumstances [1-14]. The approaches taken by most of these investigators can be divided into two categories: those using decision-directed estimates (DDE), and those using minimum mean-square error estimates (MME) of target trajectory parameters. The DDE approach labels each data point as belonging to a particular target or as random noise. The track estimates are formed by applying a conventional nonlinear filter to chains of data with a common track label. The MME approach (e.g. [1]) assigns to each data point not one but several target labels, each with an associated weight indicating the probability that the label is correct. Thus, in the MME algorithms a graduated range of possible labels is determined for each measurement, rather than the one unique label supplied by the DDE algorithms. The MME track estimates are formed by applying a weighted set of measurements to specially designed filters.

Various ways can be envisioned of feeding the available data to the above algorithms. Four of these are shown in Figure 1 below. In the first of these, all of the data from every sensor is considered simultaneously in a batch mode. In the second (hybrid) approach, data is partitioned into smaller sized sub-batches on the basis of time of occurance. The subproblems are individually solved, and patched together at their common boundaries. The third (recursive) approach is a special case of the hybrid, in which initialization occurs over a moderate sub-batch of data, and individual measurements are recursively added at each time point. Finally, the data might partitioned by sensors, the resulting subproblems solved, and the resulting estimates combined in some reasonable way. For example, tracks could be formed using the radar data alone, and then infrared data alone. Similar tracks from the individual sensors would then be combined into single (possibly more accurate) tracks. Dissimilar tracks would be regarded as real tracks seen by one sensor but missed by another.

This paper will make use of a batch implementation of the DDE approach, which leads to a large-scale combinational optimization problem. The current literature generally avoids this approach because of its formidable computational difficulties. Two avenues are available for alleviating these difficulties: (1) restatement of the problem in a more efficient computational form, and (2) increased computer speed. It has been previously shown [3-7] that this optimization problem can be stated as a 0-1 integer program:

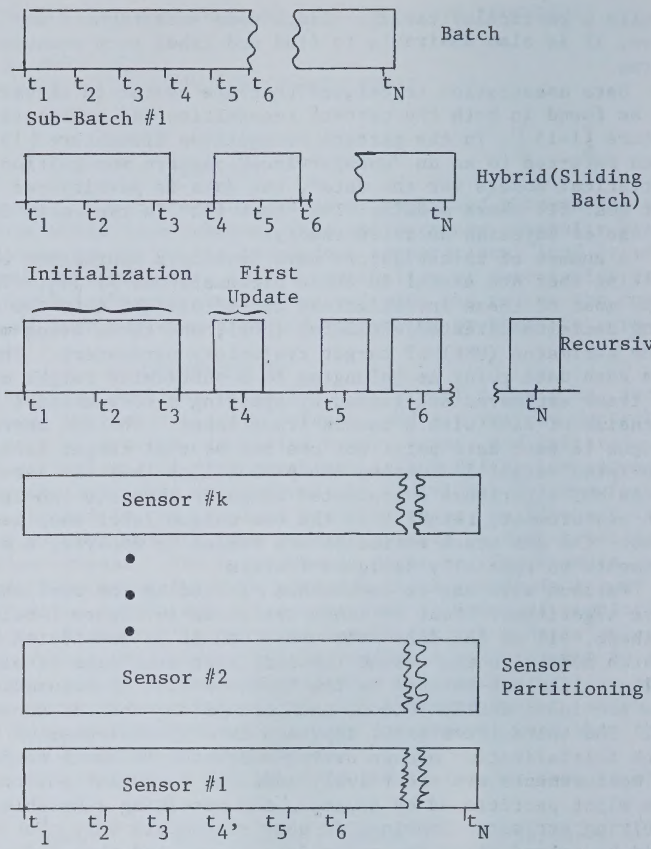
$$\min_{\tau} \; c^t \tau \tag{1}$$

$$\text{subject to} \quad A\tau \leq 1 \quad , \tag{2}$$

where τ and A are a binary vector and binary matrix respectively, and 1 is a vector of 1's. Techniques are available for solving such a problem very

---

*Such models would be derived in the present case from target motion and sensor models.



Batch
t_1 t_2 t_3 t_4 t_5 t_6 t_N
Sub-Batch #1
Hybrid(Sliding Batch)
t_1 t_2 t_3 t_4 t_5 t_6 t_N
Initialization
First Update
Recursive
t_1 t_2 t_3 t_4 t_5 t_6 t_N
Sensor #k
Sensor #2
Sensor Partitioning
Sensor #1
t_1 t_2 t_3 t_4 t_5 t_6 t_N

FIGURE 1. Organizing Data for Input to Tracking Algorithms.

efficiently[3-7].

In this paper the benefits of increased computer speed are analyzed in some detail. In particular, we examine the impact of associative (content addressable) parallel architecture [16] on the problem (1)-(2). Conventional random access memory systems require that an address in some coordinate system be specified for any word to be read or written into. Associative (or content addressable) memory systems have this capability, with the addition that any work may be addressed through its contents rather than its coordinate location. For example, given a specific bit pattern every word in memory can be simultaneously examined, and those with matching bit patterns flagged for future reference. If, in addition, parallel processing capabilities are added, then each word that has been flagged can be simultaneously operated on. Computers with the capabilities just described are often referred to as content addressable parallel processors [17].

In this paper we first will give a brief background discussion of multisensor tracking (we show that it is equivalent to problem (1)-(2)). We then compare the algorithms for this problem as they might be implemented on both a standard computer and on a content addressable parallel processor.

2. DECISION-DIRECTED APPROACH TO MULTISENSOR TRACKING

Suppose that s targets are being viewed by the various sensors of a surveillance system, and that each target is presented by a separate stochastic process with the standard Kalman form

$$x^i_{k+1} = A^i x^i_k + B^i u^i_k , \qquad \begin{array}{l} k = 0,1,\ldots,n, \\ i = 1,2,\ldots,s. \end{array} \tag{3}$$

The subscripts k denote sample times, and the superscripts i denote "target number" (an arbitrary identification label for each target). Assume that each target's position is sampled by every sensor at each time point k (we make this assumption for tutorial simplicity, since the ideas have obvious generalizations to the imperfect sensor case). If there are N sensors present, then a set of $Z_k = \{z^j_k\}^N_{j=1}$ measurements are available at each time point k. The superscripts j are arbitrary labels for the set of sN measurements available at each sample time. As we pointed out in the introduction, the main difficulty of multisensor tracking problems is the lack of a well-defined target-to-measurement correspondence $x^i_k \longleftrightarrow z^j_k$.

In order to present this difficulty in the familiar terms of linear estimation theory, we first form a big vector x_k using the $\{x^i_k\}^s_{i=1}$:

$$x_k = \begin{pmatrix} x^1_k \\ x^2_k \\ \vdots \\ x^s_k \end{pmatrix} , \tag{4}$$

and define A, B, and u so that

$$x_k = Ax_k + Bu_k , \quad k = 0, 1,\ldots,n \tag{5}$$

represents the concantenation of all the systems in (3). Similarly, let

$$z_k = \begin{pmatrix} z^1_k \\ z^2_k \\ \vdots \\ z^{sN}_k \end{pmatrix} . \tag{6}$$

At this point, in a usual estimation problem we would define a single measurement relationship. However, because of the uncertainties of multisensor tracking, we are forced here to specify a class of such relationships:

$$z_k = C^i x_k + v_k , \quad k = 1, 2,\ldots,n. \tag{7}$$

The coefficient matrices C^i come from a finite set S_c. The set S_c is obtained by finding every possible permutation of the target-to-measurement correspondences $x^i_k \longleftrightarrow z^j_k$. We assume that the a priori probability of occurance of each $C^i \epsilon S_c$ is equal.

At each sample time k, the number of elements in S_c is $(s!)^N$, so that the system described by (5), (7) leads to tremendous computational burdens. In essence, we must look at every one of the $(s!)^N$ possible systems in order to decide which one best fits the measurements z_k.

Suppose that τ represents the hypothesis that one such system is the correct one (there are $(s!)^N$ such hypotheses). Then, following a Bayesian approach, we must find the solution of

$$\max_{\tau} \; p(Z_1,Z_2,\ldots,Z_k \mid \tau) \tag{8}$$

Let λ^i be the set of measurements associated with a particular track by hypothesis τ (there are s such tracks $\{\lambda^j\}^s_{j=1}$ for each hypothesis). Assuming that each track is independent of every other track, (8) becomes

$$\max_{\tau} \; \prod^s_{j=1} p(\lambda^j \mid \tau) \quad . \tag{9}$$

Each track λ^j defines a complete set of labels for the measurements it contains. That is, each λ^j specifies to which sensor and to which track each of its data points correspond. As was done in [3-7], we can define 0-1 vectors that are equivalent to τ and λ, and convert problem (9) to the 0-1 integer program (1) - (2).

The particular problem displayed in (1)-(2) is referred to as a "set packing" problem [18]. The elements c_i of the vector c each represent the "score" for a particular track λ^i. In this case,

$$c_i = -\ln p(\lambda^i). \tag{10}$$

Any particular hypothesis τ represents a selection of such tracks, and the (negative log) likelihood of τ is just

$$c^t\tau = \sum_i c_i\tau_i = \sum_i - \ln p(\lambda^j \mid \tau) . \tag{11}$$

Since the hypotheses τ represent various permutations of the correspondences $x_k^i \longleftrightarrow z_k^j$, we can represent the constraints on τ as

$$A\tau = (\lambda^1, \lambda^2, \lambda^3, \ldots)\tau \le \begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix} \tag{12}$$

Here, λ^i is a unique bit pattern corresponding to each set of $x_k^j \longleftrightarrow z_k^j$ correspondences (for the ith track), so A is a matrix of 0's and 1's.

The reader is referred to [3-7] for a more detailed explanation of the above ideas.

3. SOLVING 0-1 INTEGER PROGRAMS

As we noted above, a brute force enumeration of every possible hypothesis τ is completely unreasonable. Several sophisticated techniques are available for (1)-(2). The implicit enumeration approach [18] will be used below. In overview, the algorithm proceeds by constructing binary solutions of the inequalities

$$c\,\tau^i < c^t\,\tau^* \tag{13}$$

$$A\tau^i \leqslant 1 , \tag{14}$$

where τ^i is to be determined and τ^* is a last best solution (initially $\tau^*=0$). If (13)-(14) cannot be solved, then τ^* is an optimal solution.

Single-Instruction Single-Data Stream Computers

Using conventional single instruction single-data stream (SISD) computers, the success of the algorithm depends on the tightness of certain necessary conditions that are imposed as the computations proceed. The matrix A is rearranged and partitioned into subjects of columns $F_1, F_2, \ldots, F_p$, where the first non-zero bit of $\lambda \epsilon F_i$ is λ_i (this gives A a staircase form). The tracks λ^j (i.e. columns of A) are reordered within each partition F_i in terms of increasing scores c_j.

The algorithm selects one element from each set F_i in turn, until either (1) any of several necessary conditions are violated, or until (2) an element (i.e. track) has been selected from each set F_i. In case (1), backtracking occurs to a previous set F_{i-1}, and the algorithm proceeds as before. In case (2), backtracking occurs to set F_{p-1}, after first setting $\tau^* = \tau^i$ where τ^i is the solution just obtained. (Recall from (13)-(14) that τ^* is the current best solution).

After an element has been chosen from F_{i-1}, a binary subproblem remains:

$$\min \sum_{j=\theta}^{r} c_j \tau_j \tag{15}$$

$$\sum_{j=\theta}^{r} \lambda^i \tau_j \leq 1 - \sum_{j=1}^{\theta-1} \lambda^j \tau_j , \tag{16}$$

where "θ" is the index of the first feasible track in F_i, and "r" is the number of tracks available to form hypotheses τ. Define

$$\phi_i = \min_{p} \left\{ c_j \middle| \lambda^j \epsilon F_i \right\}, \tag{17}$$

$$\sigma_i = \sum_{j=i} \phi_j . \tag{18}$$

Two necessary conditions to be satisfied by the track λ^q selected (from F_i) are:

$$\lambda^q \leq 1 - \sum_{j=1}^{\theta-1} \lambda^j \tau_j \tag{19}$$

and

$$c_q \leq c^t \tau^* - \sum_{j=1}^{\theta-1} c_j \tau_j - \sigma_{i+1} , \tag{20}$$

(where τ^* is the last best solution).

The above algorithm for conventional SISD computers can be seen to be a depth-first tree search. Even in this relatively efficient form, such an algorithm in general requires an "exponentially" increasing run time as problem size increases.

Content Addressable Parallel Processors

The relevant conceptual components of the content addressable parallel processor (CAPP) for this paper are shown below in Figure 1. The basic operation we are interested in is one in which a comparand is broadcast simultaneously to every one of r words stored in the content addressable memory. Each of these r words can be simultaneously compared (with the comparand), and all words meeting a specified criterion can be flagged for further use. The lowest level of such comparisons is one in which the bit setting of each word is compared to the bit setting of the comparand. If the bit patterns are the same, a tag bit for the word is set to 1 (otherwise, the tag bit remains 0). Each word with a tag bit of "1" can then be operated on in parallel. More complex comparisons can be made by logically combining a sequence of elementary comparisons.

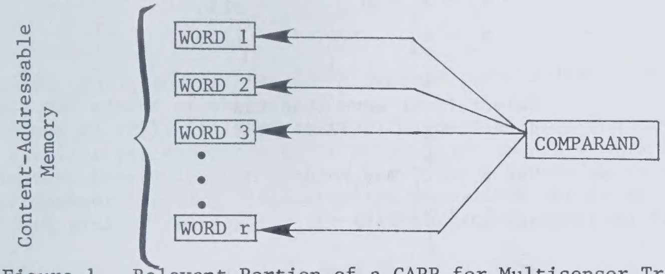


Figure 1. Relevant Portion of a CAPP for Multisensor Tracking Algorithm

The solution of the multisensor tracking problem (1)-(2) is considerably simpler using a CAPP. A rough outline of one possible algorithm is shown below in Figure 2. We assume at the outset that each pair (λ^i, c_i) is contained in one word of associative memory (for large problems several words would be needed to represent each pair (λ^i, c_i).) The comparand is then set to $(0, c^t\tau^*)$.

Every word in associative memory satisfying

$$\lambda^j \leqslant 1$$

$$c_j < c^t\tau^*$$

is flagged. The first such word (say word j_1) , is chosen, and the comparand is set to

$$(0 \oplus \lambda^{j_1} \;, \; c^t\tau^* - c_{j_1}) \;,$$

where "$\oplus$" is the inclusive "or" operator.

Every word in associative memory satisfying

$$\lambda^j \leqslant 1 - \lambda^{j_1}$$

$$c_j < c^t\tau^* - c_{j_1}$$

is flagged. The first such word (say word j_2) is chosen, and the comparand is set to

$$(0 \oplus \lambda^{j_1} \oplus \lambda^{j_2}, \; c^t\tau^* - c_{j_1} - c_{j_2}) \;.$$

These iterations continue until s tracks have been selected, at which point a complete solution has been found. This solution becomes the current best solution τ^* , and the iterations continue.

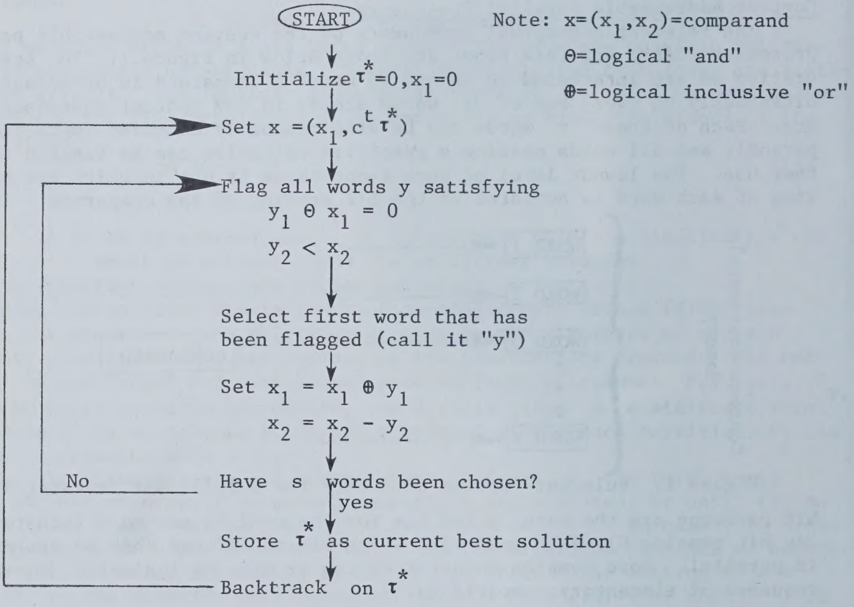

START
Note: $x=(x_1,x_2)$=comparand
Θ=logical "and"
$\oplus$=logical inclusive "or"
Initialize $\tau^*=0, x_1=0$
Set $x = (x_1, c^t\tau^*)$
Flag all words y satisfying
$y_1 \;\Theta\; x_1 = 0$
$y_2 < x_2$
Select first word that has been flagged (call it "y")
Set $x_1 = x_1 \oplus y_1$
$x_2 = x_2 - y_2$
No
Have s words been chosen?
yes
Store τ^* as current best solution
Backtrack on τ^*

FIGURE 2. CAPP Algorithm for Multisensor Tracking.

Comparison of SISD and CAPP Algorithms

The SISD algorithm presented above is a rather efficient form of implicit enumeration for the set packing problem given the architectural constraints of a conventional computer. To get an approximate idea of the computation it entails, let ϵ_i be the number of elements of each partition F_i of the columns of the constraint matrix A. Then at most, we have to examine

$$\prod_{i=1}^{p} \epsilon_i \tag{21}$$

different combinations of tracks. In general, a smaller number of track combinations is examined, since the necessary conditions that are imposed are able to discard a large number of combinations "implicitly".

Consider now just the "first pass" through the SISD algorithm, i.e. the operations required to establish the first nontrival τ^*. In order to do this, one track λ must be selected from each F_i in turn. Because of the constraints (19)-(20), several tracks in each F_i must generally be examined before a suitable candidate is found. If the expected number of discards for F_i is denoted $\hat{\epsilon}$, and this number is independent and identically distributed for all $1 \leq i \leq p$, then to establish the first τ^* requires an expected number of operations

$$(\hat{\epsilon})^s \tag{22}$$

Thus, as problem size grows, the operation count to establish the first complete solution τ^* of the SISD algorithm grows exponentially. In general, as the size of the multisensor tracking problem increases, both $\hat{\epsilon}$ and s will increase dramatically. (Recall from the previous development and [3-7] that s is the number of targets seen by the sensors).

Focusing now on the simple CAPP algorithm outlined in Figure 2, we see that the operation count to establish the first τ^* is always proportional to s.

4. SUMMARY

A comparison was made of a relatively sophisticated SISD and relatively crude CAPP algorithm for multisensor tracking. If we let T_{SISD} and T_{CAPP} be the respective times required to obtain a first complete solution τ^* to the tracking problem, we find that

$$\frac{T_{CAPP}}{T_{SISD}} \sim \frac{s}{(\hat{\epsilon})^s} , \tag{24}$$

where $\hat{\epsilon} \gg 1$, $s \gg 1$, and both $\hat{\epsilon}$ and s increase monotonically with the number of targets and number of sensors.

Based on (23), it is reasonable to conclude that the content-addressable parallel processor can offer major improvements in run speed for the combinatorial optimization problem resulting from the batch DDE approach [3-7] to multisensor tracking. Although the paper does not do so, it is easy to establish similar results for the other data organizations displayed in Figure 1.

REFERENCES

1. D.L. Alspach, "A gaussian sum approach to the multitarget identification-tracking problem," Automatica, Vol. 11, May 1975, pp. 285-296.

2. Y. Bar-Shalom, "Extension of the probabilistic data association filter to multitarget tracking," Fifth Symposium on Nonlinear Estimation Theory and Its Applications, San Diego, Calif. Sept. 1974, pp.16-21.

3. C.L. Morefield, "Solution of Multiple Choice Estimation Problems via 0-1 Integer Programming," IEEE Conf. Dec.Cont., pp. 753-754, Nov. 1974.

4. C.L. Morefield, "Efficient computational forms for Bayesian multitarget tracking," Sixth Symposium on Nonlinear Estimation and Its Applications, San Diego, California, September 1975, pp. 208-216.

5. C.L. Morefield, "Application of 0-1 Integer Programming to a Track Assembly Problem," SAMSO-TR-75-186, 16 April 1975, and IEEE Conf. Dec. Cont., December 1975, pp. 428-433.

6. C.L. Morefield, "Measurement Assignment via 0-1 Integer Programming," Aerospace Corp, TOR-0076 (6085-50)-1, 7 July 1975.

7. C.L. Morefield, "Application of Bayesian Decision Theory to Multitarget Surveillance Problems," Proceedings of NAECON '76 pp. 489-494, May 1976.

8. R.G. Sea, "Optimal Correlation of Sensor Data with Tracks in Surveillance Systems," Sixth International Conference on Systems Science, Honolulu, Hawaii, January 1973, pp. 424-426.

9. R. Singer and R.G.Sea, "New Results in Optimizing Surveillance System Tracking and Data Correlation Performance in Dense Multitarget Environments," IEEE Trans. Auto. Cont., AC-18, pp. 571-581, December 1973.

10. R.A.Singer, R.G. Sea and K.B. Housewright, "Derivation and evaluation of improved tracking filters for use in dense multitarget environments," IEEE Trans. on Info. Theory, Vol. IT-20, July 1974, pp. 423-432.

11. R.A.Singer and A.J. Kanyuck, "Computer Control of Multiple Site Track Correlation," Automatica, Vol. 7, pp. 455-463, July 1971.

12. A.J. Kanyuck and R.A. Singer, "Correlation of Multiple-Site Track Data", IEEE Trans, Aerospace Election, Syst, AES-6, pp. 180-187, 1970.

13. R.W. Sittler, "An optimal data association problem in surveillance theory," IEEE Trans. Mil. Elec. Vol. MIL-8, April 1964, pp. 125-139.

14. P. Smith and G. Buechler, "A Branching Algorithm for Discriminating and Tracking Multiple Objects," IEEE Trans. Auto. Cont., AC-20, February 1975, pp. 101-104.

15. R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, John Wiley, New York, 1973.

16. B. Parhami, "Associative Memories and Processors: An Overview and Selected Bibliography," Proc. IEEE , Vol 61, pp. 722-730, June 1973.

17. C.C.Foster, Content Addressable Parallel Processors, Van Nostrand Reinhold, New York, 1976.

REFERENCES (continued)

18. R.S. Garfinkel, G.L. Nemhauser, Integer Programming, John Wiley, New York, 1972.

LARGE SCALE LINEAR DYNAMIC PROGRAMMING
USING THE METHOD OF MULTIPLIERS

LAWRENCE E. BERMAN
The Analytic Sciences Corporation
Six Jacob Way
Reading, Massachusetts 01867

ABSTRACT

A variety of optimal control problems and dynamic multistage economic problems may be posed as large scale linear programs. A drawback to this approach is that the computational burden for solving these problems by standard techniques may be prohibitive. Recent results indicate that the method of multipliers may render these problems tractable.

1. INTRODUCTION

This paper is concerned with the solution of the following class of discrete time linear optimal control problems:

Problem 1: Find $[u(0),\ldots,u(N-1)]$ and $[x(0),\ldots,x(N)]$ which minimize

$$J = a(N)'x(N) + \sum_{i=0}^{N-1} (a(i)'x(i) + b(i)'u(i)) \tag{1}$$

subject to the system equations

$$x(i+1) = A(i)x(i) + B(i)u(i) + s(i) \quad i = 0,\ldots,N-1 \tag{2}$$

and constraints

$$C(N)x(N) \le d(N), \; C(i)x(i) + D(i)u(i) \le d(i) \; i=0,\ldots,N-1 \tag{3}$$

$x(i)$, $u(i)$, $s(i)$, $d(i)$ are real vectors. $A(i)$, $B(i)$, $C(i)$, $D(i)$ are matrices of appropriate dimension. The initial state $x(0)$ and number of stages are given. Prime denotes transposition.

Problems of this type may arise naturally from discrete time control systems or as discretizations of continuous time systems. Such problems encompass the class of problems referred to as dynamic linear programs.

It is clear that Problem 1 can be recast in the form:

Problem 2: Find $x \in R^n$ which minimizes $c'x$ subject to $\alpha_i'x \le b_i \quad i = 1,\ldots,m$

Problem 2 is the general static linear program. Once posed in this form there is an abundance of techniques (e.g., the simplex algorithm) available for its solution. The dimension of the resulting problem is proportional to N. Hence the reformulation, with the addition of slack variables, may involve hundreds of variables. Computation time is proportional to N^3.

It was demonstrated in [1], [2] that the method of multipliers [3], [4] may be employed to solve Problem (1) and utilize the dynamic problem structure. Computation time is proportional to N^2 which may result in considerable computational savings for large-scale problems.

In the remaining sections we elaborate further on the algorithm and present some computational results.

2. THE STATIC LP ALGORITHM

We consider briefly the application of the method of multipliers to the static linear program (Problem 2). The method of multipliers finds the optimal solution x^* as:

$$x^* = \lim_{k \to \infty} \left[\underset{x}{\operatorname{argmin}} \left(c'x + \frac{1}{2\mu_k} \sum_{i=1}^{m} \left[\max(0, y_k^i + \mu_k(\alpha_i' x - b_i)) \right]^2 - (y_k^i)^2 \right) \right] \tag{4}$$

$\{\mu_k\}$ is a sequence of non-decreasing positive penalty parameters. $\{y_k\}$ is a sequence of "multiplier" vectors which converge under mild assumptions [1], [2] to a Lagrange multiplier y^* if it is generated by the rule:

$$y_o = \text{any initial estimate for } y^*$$

$$y_{k+1} = \max(0, y_k^i + \mu_k(\alpha_i' x_k - b_i)) \qquad i = 1, \ldots, m \tag{5}$$

x_k represents the sequential solution of (4). The method of multipliers reduces the static linear program to a sequence of unconstrained piece-wise quadratic minimization problems. Each sub-problem may be solved in a finite number of steps by Newton or Quasi-Newton techniques. In Section 3 we discuss the implementation of a Newton's method algorithm for Problem 1.

3. THE DYNAMIC LP ALGORITHM

Problem 1 may be solved as follows:

Step 1: Select an arbitrary penalty parameter $\mu_1 > 0$. Select an initial estimate for the sequence of multiplier vectors $[y_1(0), \ldots, y_1(N)]$. Set $k = 1$.

Step 2: At iteration k solve the piece-wise quadratic unconstrained linear optimal control problem.

Problem 3: Find $[u_k(0), \ldots, u_k(N-1)]$, denoted by u_k, and $[x_k(0), \ldots, x_k(N)]$, denoted by x_k, to minimize the augmented Lagrangian

$$L(u, x, y_k, \mu_k) = f_N(x(N), y_k(N)) + \sum_{i=0}^{N-1} f_i(x(i), u(i), y_k(i)) \tag{6}$$

where

$$f_N(x(N),y_k(N)) = a(N)'x(N) + \frac{1}{2\mu_k}\{||[y_k(N)+ \mu_k(C(N)x(N)-d(N)]_+||^2 - ||y_k(N)||^2\}$$

$$\begin{aligned} f_i(x(i),u(i),y_k(i)) = {} & a(i)'x(i) + b(i)'u(i) \\ & + \frac{1}{2\mu_k}||[y_k(i)+ \mu_k(C(i)x(i)+D(i)u(i)-d(i))]_+||^2 \\ & - \frac{1}{2\mu_k}||y_k(i)||^2 \qquad i = 0,\ldots,N-1 \end{aligned} \tag{7}$$

In (7), $[\xi]_+$ denotes the vector with coordinates $\max\{0,\xi_i\}$. $||\cdot||$ is the Euclidean norm. The minimization is also subject to the system equations (2).

Problem 3 may be solved by a Newton's method of the type:

$$u_{j+1} = \begin{cases} u_j-\lambda_j\left[\nabla^2 L(u_j,x_j,y_k,\mu_k)\right]^{-1}\nabla L(u_j,x_j,y_k,\mu_k) & \text{if the inverse exists} \\ B(u_j,x_j,y_k,\mu_k) & \text{otherwise} \end{cases} \tag{8}$$

where $B(u_j,x_j,y_k,\mu_k)$ is determined by some algorithm which in the limit will generate minimizing points for convex differentiable functions, e.g., steepest descent. u_j,x_j represent the most recent estimates of u_k,x_k. $\nabla^2 L$ and ∇L represent the Hessian and gradient respectively of the augmented Lagrangian L. λ_j is a stepsize parameter which may be generated by an Armijo type stepsize rule [6]. As demonstrated in [1] only a finite number of iterations of the type (8) are required to solve Problem 3.

The Newton direction $d_j = [\nabla^2 L]^{-1}\nabla L$ need not be generated by direct computation of the Hessian. It is to be noted that if the constraints which are inactive at (u_j,x_j) are temporarily dropped, the resulting problem is purely quadratic. This may in turn be readily solved for "optimal" feedback gains and controls via the solution of a discrete Riccati equation. If the solution of this intermediate problem is denoted by $\bar{u}_j$ then the Newton direction d_j is given by

$$d_j = u_j - \bar{u}_j \tag{9}$$

Thus Problem 3 has been reduced to solving a sequence of discrete Riccatti equations. This is computationally more efficient than direct evaluation of the inverse Hessian since the matrix inversions in the solution of the Riccatti equation are low order.

A detailed description of Newton's method described above may be found in [1]. It is quite similar to techniques developed in [5] and [6].

In addition, if computation of the gradient is required the solution may be easily obtained by means of the adjoint equation. This is a standard procedure described in detail in [5].

Step 3: Update the multipliers y_k and penalty parameter μ_k by the rule:

$$y_{k+1}(N) = [y_k(N) + \mu_k(C(N)x_k(N)-d(N))]_+$$

$$y_{k+1}(i) = [y_k(i) + \mu_k(C(i)x_k(i)+D(i)u_k(i)-d(i))]_+ \qquad (10)$$

$$i = 0,\dots,N-1$$

Set $\mu_{k+1} = r\cdot\mu_k \quad r \geq 1$

As shown in [1] only a finite number of iterations of the type (10) are required for convergence.

Step 4: Check the necessary conditions for optimality. A possible convergence criterion is:

$$L(u_k,x_k,y_k,\mu_k) = L(u_{k+1},x_{k+1},y_{k+1},\mu_{k+1})$$

$$= a(N)'x_k(N)+\sum_{i=0}^{N-1}(a(i)'x_k(i)+b(i)'u_k(i)) \qquad (11)$$

Then u_k is optimal. If (11) is not satisfied set k = k+1 and return to Step 2.

In the next section the results of some computational experiments are reported.

4. COMPUTATIONAL RESULTS

A computer program was written to solve the following minimum fuel problem for the two dimensional system

$$x(i+1) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} x(i) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(i) \qquad i = 0,\dots,N-1 \qquad (12)$$

with cost functional

$$J = \sum_{i=0}^{N-1} |u(i)| \qquad (13)$$

and constraints

$$x(N) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad (14)$$

$$-1 \leq u(i) \leq 1 \qquad i = 0,\dots,N-1 \qquad (15)$$

This is a linear program if the cost functional is rewritten as

$$J = \sum_{i=0}^{N-1} u(i) + \sum_{i=0}^{N-1} z(i) \tag{16}$$

with additional constraints

$$\begin{aligned} -2u(i) &\le z(i) \\ 0 &\le z(i) \qquad i = 0,\ldots,N-1 \end{aligned} \tag{17}$$

The results of the computation for a variety of initial states and values of N are listed in Table I. Also listed is the total number of Newton iterations required for convergence. The program was written in Fortran IV and implemented on a DEC 10 system. The CPU times are listed.

TABLE I

DATA FOR DISCRETE TIME PROBLEM

N	S	x(0)	CPU TIME
10	14	(2,2)	4.40
50	90	(10,10)	2:38.06
100	135	(99.8,0)	8:34.72
200	253	(40,40)	22:18.30

N = number of stages
S = number of Newton steps
x(0) = initial state
CPU time in minutes:seconds

This technique was also tested for the two dimensional continuous time system

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= u \end{aligned} \tag{18}$$

with $x_F = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and constraints $|u(t)| \le 1$, $t \varepsilon [0,1]$. The cost function was $\int_0^1 |u(t)|dt$. The initial state was $x(0) = \begin{bmatrix} .2 \\ 0 \end{bmatrix}$.

The approach taken was to approximate the problem with an associated discrete time version. N is the number of discretization intervals. The problem was solved successively for N = 10,20,40,80,160,320,640. The starting point for each successive problem was generated from the solution of the previous approximating problem. The results are reported in Table II.

TABLE II

DATA FOR CONTINUOUS TIME PROBLEM

N	S	J
10	24	0.56000
20	11	0.55556
40	34	0.55294
80	11	0.55286
160	19	0.55282
320	23	0.55280
640	54	0.55279
* ∞		0.55279

*for N = ∞ we list the solution of the continuous time problem obtained from the maximum principle

N = number of stages
S = number of Newton steps
J = optimal value of cost functional corresponding to N

5. INTERPRETATION OF RESULTS

For the discrete time problem we note that the number of Newton steps (see Table I) required was proportional to N. Computation times were proportional to N^2. Overall computation time was, however, somewhat lengthy.

Results for the continuous time problem appear more promising. Table II indicates that successive cycles required a more or less constant and small number of steps to achieve convergence. The final discretization (N=640) represents a large scale linear program with at least 2560 variables and 5120 equality constraints. CPU time was approximately 35 minutes which is much smaller than might be predicted from Table I. Since the number of Newton steps per cycle was about constant rather than proportional to N, computation time grows linearly with N, rather than quadratically (or cubically for standard techniques). Thus an efficient version of this algorithm may be able to solve very large scale problems and permit for solution of continuous time problems with considerable precision.

6. SUMMARY

In this paper we have presented a new algorithm for the solution of linear dynamic programs. This algorithm is based upon the method of multipliers and Newton's method for linear optimal control. Its strongest area of potential application appears to be in the solution of certain continuous time optimal control problems.

7. ACKNOWLEDGEMENT

This paper is based upon results reported in the author's M.S. thesis [1] completed at the University of Illinois, under the supervision of Professor D.P. Bertsekas.

REFERENCES

1. Berman, L.E., "Linear Programming With the Method of Multipliers," M.S. Thesis, University of Illinois, Urbana, July, 1975.

2. Bertsekas, D.P., "Newton's Method for Linear Optimal Control," Prepared for the IFAC Symposium on Large Scale Systems, Udine, Italy, June, 1976.

3. Bertsekas, D.P., "On the Method of Multipliers for Convex Programming," IEEE Transactions on Automatic Control, Vol. AC-20, 1975, pp. 385-388.

4. Bertsekas, D.P., "On Penalty and Multiplier Methods for Constrained Minimization," SIAM Journal on Control and Optimization, February, 1976.

5. Dyer, P., and McReynolds, S.R., The Computation and Theory of Optimal Control, Academic Press, New York, 1970.

6. Polak, E., Computational Methods in Optimization, Academic Press, New York, 1971.

A COMPUTER PROGRAM TO FIND THE KERNEL OF A POLYNOMIAL OPERATOR*

R.R. GEJJI
Department of Electrical Engineering
University of Notre Dame, Notre Dame, Indiana

ABSTRACT

This paper presents a FORTRAN program written to solve for the kernel of a matrix of polynomials with real coefficients. It is an implementation of Sain's free modular algorithm for solving the minimal design problem of linear multivariable systems. The structure of the program is discussed, together with some features as they relate to questions of implementing the above method. An example of the use of the program to solve a design problem is included.

1. INTRODUCTION

It has been shown recently that a class of problems in linear multivariable control system design, especially those involving a requirement for proper rational solutions, can be viewed as a search for the kernel of a linear operator between free polynomial modules. When seen in this light, a typical design problem may be reduced to an equation of the type

$$[J_1(s) \;\vdots\; J_2(s)] \begin{bmatrix} N(s) \\ \hline D(s) \end{bmatrix} = [0] \tag{1}$$

where J_1 and J_2 are known matrices of polynomials, and

$$G(s) = N(s)\, D^{-1}(s) \tag{2}$$

is the matrix fraction representation chosen for the system to be designed.

In Section 2, we detail the method for solving (1), as proposed by Sain [3]. When the problem of (1) is formulated in terms of polynomials with real coefficients, the method of [3] inevitably leads to nontrivial calculations. This work is motivated by the need to mechanize these computations. Section 3 describes, in some detail, the structure of the computer program. Section 4 considers the input and output that go with the use of the program. In Section 5 we present some features of the program, seen to be relevant to problems of implementing the method of Section 2. Finally, Section 6 presents an example of the use of the program.

2. THE MDP ALGORITHM

Equation (1) can be completely solved by the method of [3], described in this section. The algorithm was originally proposed for the more general minimal design problem (MDP) formulated over polynomials in an indeterminate 'v', with coefficients from an arbitrary field. In that case, MDP reduces to an equation of the type (1), with 'v' replacing 's', the complex frequency. This is a three-part algorithm, included here for the

---

*This work was supported in part by the National Science Foundation under Grants ENG 75 22322 and GK 37285 and the National Aeronautics and Space Administration under Grant NSG 3048.

sake of completeness. The first part involves the use of unimodular column operations to calculate a basis for the kernel of a matrix of polynomials. The second part is to "reduce" this basis in a specified manner. The third part entails a search for suitable combinations of basic elements, and provides answers to questions about a linear dynamical realization for the system to be designed. The appeal of the method stems from its ability to find all possible solutions, and not merely one. We now consider the actual algorithm. For our case, T is the first matrix, $[J_1 \vdots J_2]$ in the left member of (1).

Algorithm 1

Step 1. To the n x m matrix T, adjoin an identity matrix I to form

$$\left[\begin{array}{c} T \\ \hdashline I \end{array}\right] . \tag{3}$$

Step 2. By unimodular elementary column operations, reduce (3) to the form

$$\left[\begin{array}{c:c} E_{11} & 0 \\ \hdashline E_{21} & E_{22} \end{array}\right] \tag{4}$$

where E_{11} has n rows, has no zero columns, and is in an echelon form.

Step 3. Then the columns of E_{11} are a basis for the image (Im T) of T, and the columns of E_{22} are a basis for the kernel (Ker T) of T.

Algorithm 2 is concerned with "reducing" the basis for Ker T, obtained from Algorithm 1, as will be described now. Essentially these are unimodular transformations to progressively eliminate common factors in N and D to get the minimal G of (2). The following steps can be applied to a list of elements: $m_1, m_2, \ldots, m_p$ in Ker T; for example, the basis from Algorithm 1.

Algorithm 2

Write each m_i, i = 1, 2,..., p as

$$m_i = \sum_{j=0}^{k_i} m_{i,j} \, s^j \tag{5}$$

where m_{i,k_i} is nonzero. The list $m_1, m_2, \ldots, m_p$ is said to be reduced if the matrix

$$[m_{1,k_1} \ m_{2,k_2} \cdots m_{p,k_p}] \tag{6}$$

has rank p.

Step 1. If the list $m_1, m_2, \ldots, m_p$ is reduced, stop. Otherwise, continue.

Step 2. Determine real numbers $f_1, f_2, \ldots, f_p$, not all zero, such that

$$\sum_{i=1}^{p} f_i m_{i,k_i} = 0. \tag{7}$$

Step 3. For the set of integers i having f_i nonzero, determine an i, denoted by i_{max} for which k_i is maximum, denoted by k_{max}.

Step 4. Perform the elementary column operation: Replace $m_{i_{max}}$ by

$$\sum_{i=1}^{p} f_i m_i s^{k_{max} - k_i} \tag{8}$$

Return to Step 1.

For p, an integer lying between one and the number of elements in the basis for Ker T, we can state the minimal design problem of linear multivariable systems. If p is such that

$$1 \leq p \leq r(\text{Ker } T), \tag{9}$$

find in Ker T p linearly independent elements

$$m_1, m_2, \ldots, m_p$$

which admit a linear dynamical interpretation (2) and which, when expressed as in (5) have

$$\sum_{j=1}^{p} k_j$$

a minimum. In our case, p is the number of rows in the matrix D(s) of (1). The problem is completely solved by the following.

MDP Algorithm

Step 1. Apply Algorithm 1 to determine a basis

$$b_i', \ 1 \leq i \leq b \triangleq r(\text{Ker } T) \tag{10}$$

for the kernel of T.

Step 2. Apply Algorithm 2 to the elements of (10) to form the reduced basis

$$b_i, \ 1 \leq i \leq b. \tag{11}$$

Step 3. Express the reduced basis (11) in the manner

$$b_i = \sum_{j=0}^{r_i} b_{i,j} s^j, \ b_{i,r_i} \neq 0 \tag{12}$$

for $i = 1, 2, \ldots, b$. Form the matrix

$$[b_{1,r_1} \ b_{2,r_2} \ldots b_{b,r_b}]. \tag{13}$$

If the rank of the matrix formed from the last p rows is not p, stop; MDP has no solution. Otherwise, continue.

Step 4. From the elements (13) in the reduced basis, select p elements

$$b_{i_1}, b_{i_2}, \ldots, b_{i_p}$$

with the properties

(i) the rank of the matrix formed from the last p rows of

$$[b_{i_1}, r_{i_1} \;\; b_{i_2}, r_{i_2} \; \ldots b_{i_p}, r_{i_p}]$$

is equal to p.

(ii) $\sum_{j=1}^{p} r_{i_j}$ is a minimum.

3. STRUCTURE OF THE PROGRAM

KERPO (for KERnel of a Polynomial Operator), also described in [4], is the program developed at Notre Dame to implement the method of Section 2. It is aimed at applications relating to design of multivariable control systems. In this section, and the rest of the paper, we propose to describe the features of this program, with an example of its use. The program, which requires about 700 IBM-FORTRAN source statements, has been tested on a number of typical problems. Typical CPU time required on an IBM 370/158 computer was approximately one minute. The program uses double precision arithmetic in order to reduce the difficulties encountered with rank testing and polynomial division. A PL/I version, which uses extended precision arithmetic, has been developed for the same reason. Though not documented here, this version is very similar to the one in FORTRAN. We now discuss the structure of the program.

3.1 Data Structure

The principal data structure is the COMMON data area. Figure 1 shows the FORTRAN source cards to establish this area. It should be mentioned here that the program treats polynomials in a particular fashion. Each polynomial is represented by an array containing its coefficients in order of ascending degree. Since the dimensions of these arrays are fixed, it is efficient to consider only a part of this array as meaningful data, depending on the problem size. MNC is the variable which specifies the number of positions in the array considered as meaningful data. Initially, for each polynomial, the first MNC positions must be specified, if necessary as zeros. The coefficients stored in this manner are followed by two flags stored in the same array. The first of these signifies a zero polynomial by taking the value 1.0; and is 0.0 otherwise. The second flag takes its value equal to the degree of the polynomial. The two flags are included because recognizing zero polynomials and extracting the degree of a polynomial are seen to be the core of the computation.

As an example, if MNC were 5, the polynomial $2s^2 + 1$ would be represented by the sequence of numbers

1.0, 0.0, 2.0, 0.0, 0.0, 0.0, 2.0

where the first five numbers represent polynomial coefficients, the sixth number indicates that the polynomial is not zero, and the seventh number is the degree.

```
      SUBROUTINE   READIN(NRM, JRC)
      IMPLICIT  REAL*8(A-H, O-Z)
      COMMON  COEFF(30,10,50), FR0(50), FR1(50), FR2(50),
     *        FR3(50), FR4(50), FR5(50), FR6(50), FR7(50)
      COMMON  XPARM, EPS, NRO, NCO, NZFL, MACE
      COMMON  MNC, IDEG, NCALLS, INDBUS
```

Figure 1. COMMON Area

We now describe in some detail the items in the COMMON area and their role.

NRM. The number of rows in the original matrix of polynomials. Not actually in the COMMON area, but serves to define NRO.

NCO. The number of columns.

NRO. (=NRM+NCO). The number of rows in the augmented matrix of Algorithm 1.

MNC. Estimated maximum number of coefficients encountered in a single polynomial during computation, that is, the maximum degree plus one.

NZFL. Position of the zero flag. (=MNC+1).

MACE. Position for storing the degree of the polynomials. (MACE = NZFL+1).

COEFF(I,J,K). The three-dimensional array where the polynomial matrix is stored. I and J indicate the row and the column, respectively, of the polynomial in the matrix and K denotes the position of the desired coefficient. Thus, for the second degree coefficient, K = 3.

FR0 thru FR7. Eight floating point arrays, each acting as a register capable of storing a polynomial. These are used for transferring polynomials between subroutines that do the polynomial arithmetic.

XPARM. User selected parameter which sets the threshold for loss of significant digits. If n numbers are being added to form a sum, the sum is compared with XPARM times the largest of the absolute values of the numbers. If the sum is smaller in absolute value, then it is set equal to zero. The default value is 1.E-10, which corresponds to 10 digits being considered significant.

For example, if we add 1.000 and -0.999, the question is whether or not the sum (0.001) is significant. If XPARM is set to 10^{-3} or greater, it will be considered nonsignificant.

EPS. In determining the degree of a polynomial, all high order coefficients with an absolute value less than this number are neglected. Thus, the degree is determined by the first leading coefficient having an absolute value greater than EPS. The default value is 1.E-40.

NCALLS. This variable keeps track of the number of times a subroutine names TEST is called and is of use in program tracing.

IDBUG. If this value is specified as one, all results of intermediate column operations during Algorithm 1 are printed. Default value is zero.

3.2 Main Program Flow

The mainline program implements in sequential order the Algorithm 1, Algorithm 2, and (Algorithm 3) the latter portions of the MDP Algorithm. For support of this computation, there are thirteen subroutines.

3.3 Polynomial Arithmetic Routines

These consist of the following five:

(i) SUBTR (A,B,DIFFP)

(ii) ADD (A,B,SUMP)
(iii) MULT (A,B,PRODP)
(iv) DIVID (A,B,QUOTP,REMAP)
(v) DEGRE (A,MM)

They are written for purposes made obvious by their names. Each subroutine performs the required operations on polynomials in any of the eight floating register arrays as specified. Results are similarly returned through a different register. There are some exceptions, and one is that FR0 and FR1 may not be used as arguments in calling DIVID, because these are used by DIVID for other purposes. Also, the DIVID routine assumes a monic divisor. The DEGRE subroutine extracts the degree by examining the coefficients. It then sets the two flags in the array accordingly.

These subroutines have some special features. Unlike the usual SSP (Scientific Subroutine Package) subroutines, these permit the dimension (subscript range) of the array holding the polynomial to be independent of the degree. They incorporate certain options provided by the program, which allow the user to choose accuracy levels and so forth. Finally, the subroutines have been written to handle, and also to provide as results, polynomials that are compatible with the representation chosen.

3.4 Polynomial Transfer Routines

The following subroutines are used to transfer polynomials back and forth between the COEFF array and the registers used for actual arithmetic purposes.

(i) STORM (IA,IB,ADDR)
(ii) XLFR (I1,I2,ADDR)
(iii) XLFA (I1,I2,I3,I4)
(iv) XMON (LX1,LX2)

The first copies a polynomial from a specified location of the polynomial matrix to a specified register array. The second subroutine does just the opposite. XLFA is used to copy a polynomial in a specified location of the polynomial matrix into another location. Lastly, XMON is used to divide by a suitable real number, all the polynomials in the column indicated by LX2, so that the polynomial at location (LX1,LX2) becomes monic. The column operation is carried over the first NR0 rows, which correspond to the augmented matrix of Algorithm 1. As will be seen later, these subroutines offer capability of column operation manipulations needed during Algorithm 1.

3.5 Program Self-Check and Tracing Routines

(i) FAULT: This routine causes the printing of a specified column of the polynomial matrix.

(ii) TEST. This subroutine is used to verify the computational success of Algorithm 1. Essentially, we use the fact that the lower portion of the augmented matrix represents a transformation, which when applied to the original matrix, yields the final form of the upper portion of the augmented matrix. For this purpose, the program uses a portion of the COEFF array to store a copy of the original matrix. If any numerical discrepancy is detected, suitable messages are printed.

(iii) WOLF. This subroutine has a function similar to the previous one. At the end of Algorithm 2, this subroutine checks to see if the elements of the reduced basis are in the kernel of the original matrix.

3.6 Rank Finder

A subroutine RANK has been written to use in testing the full rank condition during Algorithm 3.

4. INPUT AND OUTPUT

Input to the program is through a user written subroutine READIN. The COMMON area must be correctly established. Problem parameters must be initialized by the user. The polynomial matrix must be set to $[J_1 \quad J_2]$. COEFF (I,J,K) must be specified for all: $1 \leq I \leq NRM$, $1 \leq J \leq NCO$, $1 \leq K \leq MACE$. The program can handle a 10 x 10 matrix, and MNC must not exceed 48.

To set correctly the flags for all the polynomials in the matrix, the user may invoke the DEGRE subroutine. There are two parameters for the subroutine READIN. One is NRM, which is self-explanatory. The other parameter JRC is set equal to MNC.

The program, when it begins executing Algorithm 3, reads a single integer value. This must equal the desired number of rows in D(s) of (1).

Initial output is the starting input matrix. Unless IDBUG is specified as one, Algorithm 1 is completed without any further output. This is followed by the augmented matrix as it appears after completion of Algorithm 1. After this comes the output generated by the TEST subroutine, if any.

After going through Algrithm 2, the reduced basis is printed. The saved copy of the original input matrix is now printed for verification. This is followed by any output due to the WOLF routine.

Last, if there exists a minimal linear dynamical interpretation, it is printed. All the selected combinations that offer a minimal interpretation are indicated. The order of the minimal system is calculable from the reduced basis elements. This number is also printed.

5. DESCRIPTION OF MAINLINE SEGMENT

Figure 2 shows a flowchart of the program. In Algorithm 1, the condition of unimodular transformations demands that only elementary column operations be used.

(i) Interchange two columns.
(ii) Multiply a column by a polynomial and add to another column.
(iii) Multiply a column by a non-zero real number.

The strategy used here is to determine a suitable pivotal position in each row, as dictated by the requirement of the echelon form of E_{11} in (4). A type (i) transformation can be used to move a lowest degree element in the row to the pivotal position. Next, a type (ii) transformation, in conjunction with Euclid's division algorithm, can be used to reduce all succeeding polynomials in the row to a degree less than the pivotal polynomial degree. A little thought will show that, if repeated a sufficient number of times, the last two steps will yield the required form of Algorithm 1.

Type (iii) transformations are used to make the pivotal polynomial monic before performing division. This also helps in keeping the numbers in a useful range through a manner of normalization.

In Algorithm 2, the crucial steps are steps 1 and 2. It can be shown [4], that the question of Step 1 can be answered, while simultaneously providing a solution to the equation in Step 2, if needed. This is done by formulating an equivalent kernel problem involving a matrix of real numbers. The method of solution is analogous to that of Algorithm 1.

In Algorithm 3, the problem is that of methodically selecting different combinations of basis elements. For this, an array of integers is used, the elements of which are actually pointers to the latest combination of elements that is generated. The numbers in the pointer array are modified methodically, to generate all possible combinations of the required number of linearly independent basis elements. In this stage, each newly generated combination needs to be tested for a full-rank condition.

6. A DESIGN EXAMPLE

We have selected an example from Wang's work [5]. The problem is finding a realization requiring a minimum number of integrators for the transfer function G(s),

$$G(s) = \frac{M(s)}{d(s)} \tag{14}$$

where

$$M(s) = \begin{bmatrix} -3s^3 + 10s + 4 & s^3 - 3s - 1 & s + 1 \\ s^2 - 2s & s & s^2 + s \end{bmatrix} \tag{15}$$

and

$$d(s) = s^4 + s^3 - 3s^2 - 5s - 2. \tag{16}$$

Next, we let

$$G(s) = N(s)\, D^{-1}(s) = \frac{M(s)}{d(s)} \tag{17}$$

and obtain an equation of the type (1).

$$\left[dI \;\vdots\; -M \right] \left[\begin{array}{c} N \\ \hdashline D \end{array}\right] = [0] \tag{18}$$

Now, the 2 x 5 matrix [dI -M] is input to the kernel calculating software. The following solution was obtained with a value of 9 for MNC and 10^{-4} for XPARM. From the program we obtain as a reduced basis for Ker T the elements

$$\begin{bmatrix} 0 \\ -s \\ -.33s-.33 \\ 0 \\ -s^3+3.33s+1.33 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 0.33 \\ 1 \\ -.33 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ -.33s-.33 \\ 0 \\ .33s-.67 \end{bmatrix}$$

Finally, it is verified that the above reduced basis, when all three of its columns are combined, offers a minimal dynamical interpretation. The minimal order is seen to be four. Now we can write

$$G(s) = \begin{bmatrix} 0 & 0 & 1 \\ -s & 0 & 0 \end{bmatrix} \begin{bmatrix} -.33s-.33 & .33 & -.33s-.33 \\ 0 & 1 & 0 \\ -s^3+3.33s+1.33 & -.33 & .33s-.67 \end{bmatrix}^{-1} \tag{19}$$

Having obtained G(s) as in (19), it is simple to obtain, because of the properties of the reduced basis, a state space realization for G(s). This could be done, for example, by the method of [5].

7. CONCLUSION

We feel that polynomial techniques have rich potential application in the area of multivariable control system design. Research in the area is hampered by non-availability of computational aids and necessitates the researcher being bogged down in tedious calculation. The literature, for example [1], therefore contains only simple examples suitable for hand calculation. To this author's knowledge, few attempts have been made to apply these methods to a significant problem. The program presented here should prove an aid in this regard. Elliot [2] has also reported software to do similar calculations using a different approach.

The software reported here had good success with the kind of problems that have so far appeared in the literature. Numerical roundoff error appears to be the biggest problem encountered when attempts are made to apply these techniques to significant problems [4]. These attempts indicate that roundoff error introduced during implementation of the division algorithm is responsible for the most severe numerical problems. These problems could be tackled on two fronts. One would be the development of new algorithms which are less prone to numerical instability. Second, sophisticated support software offering higher precision arithmetic can be used.

8. ACKNOWLEDGEMENT

Valuable suggestions from Professor M. K. Sain, of the University of Notre Dame, are gratefully acknowledged.

9. REFERENCES

[1] Wang, S. H. and Davison, E. J., "A Minimization Algorithm for the Design of Linear Multivariable Systems," IEEE Transactions on Automatic Control, Vol. AC-18, pp. 220-225, 1973.

[2] Elliott, H., "Implementation of Computer Algorithms Related to Multivariable Systems Theory," Division of Engineering Technical Report No. NSF-Eng. 73-03846A0I/3, Brown University, Providence, Rhode Island, May 1975.

[3] Sain, M. K., "A Free Modular Algorithm for Minimal Design of Linear Multivariable Systems," Preprints Sixth IFAC World Congress, pp. 9.1-1 --9.1-7, Boston, August 1975.

[4] Gejji, R. R. and Sain, M. K., "Polynomial Techniques Applied to Multivariable Control of Jet Engines," Technical Report No. EE-761, Dept. of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana, March 1976.

[5] Wang, S. H., "Design of Linear Multivariable Systems," Electronics Research Laboratory, University of California, Memorandum No. ERL-M309, Berkeley, October 1971.

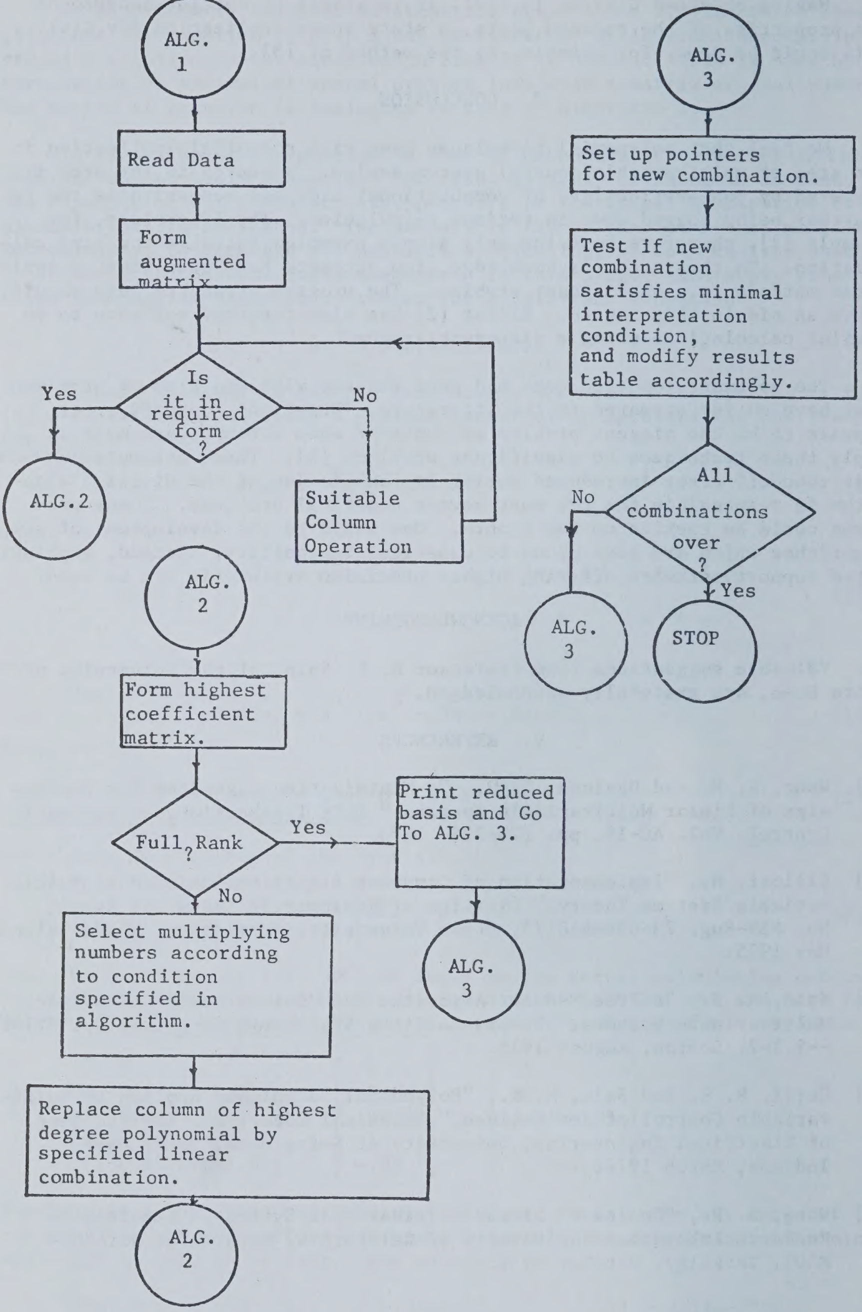

ALG. 1
Read Data
Form augmented matrix.
Is it in required form ?
Yes
No
ALG. 2
Suitable Column Operation
ALG. 2
Form highest coefficient matrix.
Full Rank ?
Yes
No
Print reduced basis and Go To ALG. 3.
ALG. 3
Select multiplying numbers according to condition specified in algorithm.
Replace column of highest degree polynomial by specified linear combination.
ALG. 2
ALG. 3
Set up pointers for new combination.
Test if new combination satisfies minimal interpretation condition, and modify results table accordingly.
All combinations over ?
No
Yes
ALG. 3
STOP

Figure 2. Outline of Main Program

ANALYSIS AND SYNTHESIS OF DYNAMIC SYSTEMS VIA BLOCK PULSE FUNCTIONS

P. SANNUTI
Department of Electrical Engineering, Rutgers University,
New Brunswick, NJ 08903

ABSTRACT

This paper presents a method of numerically integrating a system of differential equations based on an idea of orthogonal approximation of functions. Here block pulse functions are chosen as the orhtogonal set. The method gives piecewise constant solutions with minimal mean square error and is computationally similar to the familiar trapezoidal rule of integration. Design of piecewise constant controls [1] or feedback gains for dynamic systems can be simplified following this approach.

DEVELOPMENT

Consider a set of block pulse functions over a unit interval (0,1], $\phi_i(t)=1$ for $(i-1)|m<t\leq i|m$ and is zero elsewhere. Any integrable function f(t) over (0,1] can be expanded as $f(t)\simeq a_1\phi_1(t)+\cdots+a_m\phi_m(t)$ where a_i denotes the average value of f(t) over $(i-1)|m<t\leq i|m$. Denoting $\phi(t)$ as an m-vector with $\phi_i(t)$ as its i-th element, the integral of $\phi(t)$ can then be expanded into $\phi(t)$,

$$\int_0^t \phi(t)dt \simeq P\phi(t) \tag{1}$$

where P is an upper triangular matrix with all main diagonal elements equal to 1/2m and all other elements above the main diagonal equal to 1/m. Now consider the system,

$$\dot{X} = AX + BU \ , \ X(o) = X_o . \tag{2}$$

Expand the variables $\dot{X}$, X and BU into block pulse functions,

$$\dot{X} = C\phi(t) \ , \ X(t) = D\phi(t) \ , \ BU(t) = E\phi(t). \tag{3}$$

Our intention is to calculate the matricies C and D. Note the matrix E is known for a given input U(t). Inserting (3) in the identity

$\int_0^t \dot{X}\,dt = X(t) - X_o$, we get

$$[CP]_i = D_i - X_o \ , \ i = 1 \text{ to } m, \tag{4}$$

where C_i, D_i and $[CP]_i$ are respectively in i-th columns of C, D and CP. Noting the special form of the matrix P, we have

$$D_1 = (1/2m)C_1 + X_o, \ \ D_{i+1} = D_i + (1/2m)(C_{i+1} + C_i) \text{ for } i \geq 1. \tag{5}$$

To obtain another relationship between C and D insert (3) in (2),

$$C_i = AD_i + E_i \text{ for all } i \geq 1. \tag{6}$$

Finally combining (5) and (6), we have

$$D_1 = \alpha X_o + (\alpha/2m)E_1, \ \ D_{i+1} = \beta D_i + (\alpha/2m)(E_{i+1}+E_i) \text{ for } i \geq 1, \tag{7}$$

where $\alpha = (I-A/2m)^{-1}$ and $\beta = \alpha(I+A/2m)$. Note that (7) represents the familiar trapezoidal rule of integrating a system of differential equations.

1. C.F. Chen and C.H. Hsiao, "Design of Piecewise Constant Gains for Optimal Control via Walsh Functions," IEEE Trans. on A.C., Oct. 1975.

A MULTI-MICROPROCESSOR ARCHITECTURE FOR SOR-TYPE ALGORITHMS

GERARD G.L. MEYER
Electrical Engineering Department
North Carolina State University
Raleigh, North Carolina 27607

ABSTRACT

An approach for solving turbulence problems in fluid flows requires that the non-linear Poisson equation

(1) $\nabla^2(a) = f(a)$

be solved repeatedly. The large data base associated with such problems together with the inherent difficulty of solving equation (1) require computers with large primary memory and hight parallelism of computation. Instead of using specially designed computers (SIMD or MIMD) we propose to use a network of general purpose computing elements. The solution of (1) on such a network requires the development of a non negligeable amount of software. In this paper, we only discuss SOR-type algorithms for solving the Poisson equation

(2) $\nabla^2(a) = b$

with Dirichlet's boundary conditions and partitioning schemes for implementing these methods on a network of computing elements. The importance of solving (2) efficiently lies in that algorithms for solving (1) require the repeated solution or approximated solution of (2).

INTRODUCTION

This paper is devoted to the analysis of the implementation of a class of iterative algorithms for solving the Poisson equation on a network of computing elements. It is assumed that each element possesses the capabilities of a general purpose computer, i.e., capability of storing data and program, capability of performing computations on the stored data according to the stored program and finally capability of communicating with other computing elements of the network. The types of networks we are interested in are therefore special purpose networks of general purpose computing elements. Although the specific computing power of the computing elements is of no importance as far as the results of this paper are concerned, we nevertheless have in mind computing elements which are either minicomputers or microcomputers.

Our approach follows Lehman [1], Rosenfeld [2] and McGill and Steinhoff [3]. We examine the problem of solving the Poisson equation to find the macroparallelism rather than find parallelism at the instruction level [4]. The behavior of the resulting network is therefore asynchronous with a light coupling due to the exchange of data between the elements of the network.

NUMERICAL SCHEMES

There exist many methods for solving elliptic Poisson equations. A choice has to be made as to which methods are most appropriate for implementation on a network of computing elements.

In this paper, we consider only the two-dimensional Poisson equation

$$(3) \qquad \frac{\partial^2 a}{\partial x^2} + \frac{\partial^2 b}{\partial y^2} = b$$

with Dirichlet's boundary conditions, i.e., the values of a are known on the boundary. Equation (3) is discretized using a second order difference scheme which yields a "five point formula"

$$(4) \qquad \frac{a_{i+1;j} - 2a_{i;j} + a_{i-1;j}}{\Delta x^2_{i;j}} + \frac{a_{i;j+1} - 2a_{i;j} + a_{i;j-1}}{\Delta y^2_{i;j}} = b_{i;j}$$

The numerical methods for solving equation (4) may be classified into direct and iterative methods [5]. We will not investigate direct methods for the following reasons:

(i) they are limited to specific shapes;
(ii) they are limited to specific boundary conditions;
(iii) error propagation is difficult to control.

Iterative methods are either explicit or implicit. Implicit methods, i.e., line SOR, alternating directions implicit methods,...etc., require the solution of implicit systems of equation at each iteration. It is not clear if the inner iterations required to solve these systems of equation may be implemented on networks possessing a structure compatible with the structure necessary for the outer iterations. This is a question which we leave open for further studies. Explicit methods, on the other hand, do not present us with such problems: there are no inner iterations to take into account. Also it is known that explicit iterative methods which use computed points as soon as possible have a better computational behavior than methods who do not, i.e., Gauss-Seidel types methods are "better" than Jacobi-type methods [5]. For all these reasons we have chosen to study the implementation of SOR and SOR-type methods on networks of computing elements.

SOR AND SOR-TYPE METHODS

Let S_{IP} be the set of pairs (i;j) such that the mesh points (i;j) are interior points of the domain of integration and let S_{BP} be the set of pairs (i;j) such that the mesh points (i;j) are the boundary points of the domain. We assume Dirichlet's boundary conditions, i.e., the values of $a_{i;j}$ are known for all (i;j) in S_{BP}. Given initial guesses $a^0_{i;j}$ for all (i;j) in S_{IP}, SOR-type iterative methods generate successive approximations for $a_{i;j}$ by using formulas of the following form:

$$(5) \quad a^{k+1}_{i;j} = f(a^k_{i;j};a^k_{i+1;j};a^{k+1}_{i-1;j};a^k_{i;j+1};a^{k+1}_{i;j-1};i;j;k).$$

Actually, we do not need to store the values $a^k_{i;j}$ and $a^{k+1}_{i;j}$ for all (i;j) in S_{IP} provided that the mesh points are scanned in a manner compatible with equation (5). To emphasize this point, we redefine SOR-type methods using scanning maps, i.e., one-to-one maps from the set $\{1,2,3,\ldots,|S_{IP}|\}$ onto the set S_{IP}, where $|S_{IP}|$ denotes the cardinality of S_{IP}. The algorithm uses a sequence of such maps together with a stop rule which is not explicited but which may depend on k or the precision of the iterates at iteration k.

Algorithm 1: Let $s_1(.), s_2(.), \ldots$ be a sequence of scanning maps and let $a_{i;j}$ be given for all $(i;j)$ in $S_{IP} \cup S_{BP}$.

```
Step 0: Set $k = 1$.
Step 1: Set $t = 1$.
Step 2: Set $(i;j) = s_k(t)$.
Step 3: Set $a_{i;j} = f(a_{i;j}; a_{i+1;j}; a_{i-1;j}; a_{i;j+1}; a_{i;j-1}; i;j;k)$.
Step 4: Set $t = t+1$.
Step 5: If $t \leq |S_{IP}|$, go to Step 2; else go to Step 6.
Step 6: If stop condition is satisfied, stop; else, set
        $k = k+1$ and go to Step 1.
```

Definition: A scanning map $s(.)$ is said to be a North-East scanning map (NE-map) if and only if $s^{-1}(i_1;j_1) \leq s^{-1}(i_2;j_2)$ for all pairs $(i_1;j_1)$ and $(i_2;j_2)$ in S_{IP} which satisfy $i_1 \leq i_2$ and $j_1 \leq j_2$.

Lemma: If the maps $s_k(.)$, $k = 1,2,\ldots$ are NE-maps, then algorithm (1) is a SOR-type iterative method in which the updating formula is of the form of equation 5.

Notes: (i) The symbol = in Step 3 of algorithm (1) must be interpreted as an assignement, i.e., the right hand side is computed and the result assigned to the variable $a_{i;j}$.

(ii) There exist other formulas for updating the values of $a_{i;j}$, for example

$$(6) \quad a^{k+1}_{i;j} = f(a^{k}_{i;j}; a^{k}_{i+1;j}; a^{k+1}_{i-1;j}; a^{k+1}_{i;j+1}; a^{k}_{i;j-1}; i;j;k).$$

In this case, we would have to use algorithm (1) with South-East scanning maps, i.e., maps $s(.)$ such that $s^{-1}(i_1;j_1) \leq s^{-1}(i_2;j_2)$ for all pairs $(i_1;j_1)$ and $(i_2;j_2)$ in S_{IP} satisfying $i_2 \leq i_1$ and $j_1 \leq j_2$. It is clear that there exist also South-West and North-West scanning maps. When algorithm (1) is used with such maps, the $a_{i;j}$ are updated using formulas which differ from equation (5) and equation (6) in the way the superscript k+1 appears.

(iii) There exist scanning maps which do not correspond to any formula. The scanning of the mesh points which result from the use of such maps will be called non-uniform scanning and will not be discussed in this paper.

(iv) A convenient way of depicting a scanning map is to associate with every mesh point the value of its inverse at that particular mesh point. Figure 1 contains examples of various scanning maps, both uniform and non-uniform represented in this fashion.

PARTITIONING SCHEMES

There are usually many ways of implementing a serial algorithm on a network of N computing elements $CE^1, CE^2, \ldots, CE^N$. Some implementations are such that the iterates produced by the serial algorithm are not identical to the iterates produced by the algorithm implemented on N computing elements. This introduces a new difficulty in the discussion of parallel implementation. To remove it, we restrict ourselves in this paper to implementations for which such differences do not exist.

Definition: We say that an implementation of a serial algorithm on a network of computing elements is a T-implementation iff the iterates produced by the serial algorithm and its implementation on the network of computing elements are identical.

The only difference between a serial algorithm and its T-implementation resides in the time it takes to obtain the iterates and not the value of the iterates themselves. T-implementation are by no means the rule. For example, page 10 and 11 of [6], Stone discusses two implementations of the Gauss-Seidel algorithm on the ILLIAC IV. The algorithm considered is therefore a special case of algorithm (1) with NE-scanning maps and neither one of the implementations presented is a T-implementation: the first one corresponds to an algorithm half-way between the Jacobi and the Gauss-Seidel algorithm and the second one corresponds to an implementation of algorithm (1) with the non-uniform scanning given in Figure (1).

To implement algorithm (1) on a network of N computing elements, we need to partition the set S_{IP} into N blocks S_{IP}^1, $S_{IP}^2, \ldots, S_{IP}^N$ and then to assign each block of interior mesh point to a computing element. Obviously, the efficiency of the network depends on the partitioning scheme used. In the case under study we have several requirements which must be satisfied in order to have a "good" partitioning scheme. We now discuss in some detail these requirements.

Number of Communication Links

The computing element CE^n is assigned the block of interior mesh point S_{IP}^n. It must communicate with all computing elements CE^m which contain a block S_{IP}^m having a common boundary with S_{IP}^n. It is therefore clear that the number of bidirectional communication links N_{CL} depends directly on the partitioning scheme. Communication links are expensive and introduce a non-negligeable amount of complexity in the hardware. We must try to keep N_{CL} to a minimum. In fact, we will restrict ourselves to partitioning schemes which have the property that every computing element is linked to at most two other computing elements, i.e., partitioning schemes which are such that $N_{CL} \leq N$.

Computational Efficiency

Let $T_{k;1}$ be the time necessary to scan k times the set S_{IP} with algorithm (1) implemented on one computing element and let $T_{k;N}$ be the time necessary to scan k times the set S_{IP} with a T-implementation of algorithm (1) on a network of N computing elements. We define the computational efficiency $E_{k;N}$ of the T-implementation to be

$$E_{k;N} = T_{k;1} / (N.T_{k;N}). \tag{7}$$

The computational efficiency is always smaller than or equal to one. When it is equal to one, then by using the network of N computing elements, we obtain our results N times faster than if we use one computing element. The search for partitioning schemes which are optimal with respect to $E_{k;N}$ is difficult because of the complexity of the quantities $T_{k;1}$ and $T_{k;N}$. The problem is simplified by introducing the conceptual computational efficiency $E_{k;N}^*$. Let $T_{k;1}^*$ and $T_{k;N}^*$ be the time necessary to scan k times the set S_{IP} using one and N computing elements respectively, assuming that (i) the time necessary to transfer information from one computing element to the other is negligeable and (ii) the time necessary to process each interior mesh point is constant and equal to τ. We define $E_{k;N}^*$ to be

$$E_{k;N}^* = T_{k;1}^* / (N.T_{k;N}^*). \tag{8}$$

We note that if the computing elements can process and transfer information simultaneously, then $E^*_{k;N}$ will be a good approximation of $E_{k;N}$.

Data Base Storage Efficiency

It is not necessarily true that the amount of memory required for storing the data when one computing element is used is equal to the amount of memory required for storing the data when N computing elements are used. The particular manner in which algorithm (1) is implemented on the network may require that some of the data be duplicated. At a given instant of time, the same information may be present in more than one computing element. Such an occurence may be due to either the implementation of the algorithm on the network or it may be the result of a deliberate effort to decrease timing interferences between the various elements. In order to characterize the possible inefficiencies of the network concerning the storage of the data, we define the data base storage efficiency E_{DBS} to be the ratio of the amount of memory to store the data when one computing element is used to the total amount of memory required to store the data when algorithm (1) is T-implemented on a network of N computing element.

PARTITIONING ALGORITHM

A partitioning algorithm must be efficient with respect to the three criteria discussed in the preceding section. The resulting implementation should be a T-implementation, N_{CL} should be smaller than or equal to N, $E_{k;N}$ or $E^*_{k;N}$ should be as close to one as possible for all k and E_{DBS} should be as close to one as possible also. In addition, the algorithm must generate easily described boundaries between the various regions and the complexity of the algorithm itself should be small. It does not make any sense to have a partitioning algorithm which takes longer to generate the partition of S_{IP} than it takes to solve the Poisson equation on one computing element. We now present a partitioning algorithm which achieves a good compromise between the more or less contradictory requirements just given.

Algorithm 2: Let N be the number of computing elements in the network.

Step 0: Let $q_0 = \min \{j-1 | (i;j) \in S_{IP}\}$, let $g_0 = 0$ and let $k = 1$.

Step 1: Let q_k be the smallest integer satisfying
$$|\{(i;j) | q_{k-1} < j \leq q_k\}| \geq |S_{IP}|/N.$$

Step 2: Let $S_{IP}^k = \{(i;j) | q_{k-1} < j \leq q_k\}$..

Step 3: Let $g_k = g_{k-1} + |S_{IP}^k|$.

Step 4: If $|S_{IP}| - g_k = 0$, stop; If $0 < |S_{IP}| - g_k < |S_{IP}|/N$ let $S_{IP}^{k+1} = \{(i;j) | q_k < j\}$ and stop; else, set $k = k+1$ and go to Step 1.

Algorithm 2 is sub-optimal in the sense that $E_{k;N}^*$ is usually smaller than one but it has nevertheless many advantages:

(i) the algorithm generates the partition rapidly and is reasonably simple to program;

(ii) the boundaries between the various regions are easy to characterize;

(iii) the number of communication links N_{CL} is equal to N-1 at most;

(iv) it can be shown that if $|S_{IP}|$ is an exact multiple of N, then $E_{k;N}^*$ is monotonocally increasing with k , converges to one as k goes to $+\infty$ and that $E_{k;N}^* \geq k/(k+(N-1))$ for all k;

(v) $E_{k;N}^* \geq k/(k\xi + N - \xi)$ for all k, where $\xi = N\alpha/|S_{IP}|$ and $\alpha = \max \{|S_{IP}^k| \mid k = 1,2,\ldots,N\}$

We note that (iii) implies that algorithm 2 is asymptotically optimal provided that the partition is uniform. It is possible to improve the performance of algorithm 2 by considering not only North-East scanning maps, but North-West, South-East and South-West maps as well. We apply the algorithm to the region four times, each time rotating the region by $\pi/4$, then evaluating $E_{k;N}^*$ in all four cases and picking the best orientation.

The criteria which we used in this paper are different from the ones which are more commonly proposed in the literature, namely the efficiency (ratio of number of computations actually performed to number of computations possible in a given time) and the speed up ratio (ratio of the computational time on a serial computer using the best serial algorithm to computational time on the network of computing element using the best algorithm together with the best implementation). We feel that these two criteria are not well adapted to our problem. The first one does not seem very meaningful in a multi-computing element environment and the second one involves too many variables and unknowns.

The computational performance of a T-implementation of algorithm (1) on a network of N computing elements using a partition given by algorithm 2 is measured by $E_{k;N}$. We only have results concerning $E^*_{k;N}$. The lower bounds we have on $E^*_{k;N}$ are very pessimistic in general, although they may be reached in some pathological cases. Work is in progress to obtain better estimates for the lower bounds for $E^*_{k;N}$ in the case of convex domains and to relate $E_{k;N}$ with $E^*_{k;N}$. In order to get an idea of the type of results which may be obtained we now present an example in which the integration domain is simple.

Example: Assume that the domain of integration is a square, the mesh consists of 40,000 points arranged on a 200 × 200 matrix and the network consists of 10 computing elements. Then:

$N_{CL} = 9;$

$E^*_{k;10} \geq k/(k+0.045)$ for all $k = 1,2,\ldots$

$E_{DBS} = 0.95.$

REFERENCES

[1] M.Lehman,"A Survey of Problems and Preliminary Results Concerning Parallel Processing and Parallel Processors", Proc. IEEE, Vol. 54, No. 12 (1966), pp. 1889-1901.

[2] J.L.Rosenfeld,"A Case Study in Programming for Parallel Processors",Comm. ACM, Vol. 12, No. 12 (1969), pp. 645-655.

[3] R.McGill and J.Steinhoff,"A Multimicroprocessor Approach to Numerical Analysis: An Application to Gamming Problems", Proc. Symp. on Computer Architecture, (1975).

[4] W.L.Hatcher, F.M.Brasch, Jr. and J.E. Van Ness,"A Feasibility Study for the Solution of Transient Stability Problems by Multiprocessor Structures", IEEE Power Apparatus and Systems, Winter Power Meeting, September 1 1976.

[5] P.T.Roache, Computational Fluid Dynamics, Hermosa Publishers, P.O. Box 8172, Albuquerque, NM 87108.

[6] H.S.Stone,"Problems of Parallel Computation" in Complexity of Sequential and Parallel Numerical Algorithms, J.F. Traub Ed., Academic Press,1973.

```
*   *   *   *   *   *   *
*  16  17  18  19  20   *
*  11  12  13  14  15   *
*   6   7   8   9  10   *
*   1   2   3   4   5   *
*   *   *   *   *   *   *
```

North-East scanning

```
*   *   *   *   *   *   *
*  12  14  15  19  20   *
*   5   7   9  13  18   *
*   2   4   8  11  17   *
*   1   3   6  10  16   *
*   *   *   *   *   *   *
```

North-East scanning

```
*   *   *   *   *   *   *
*   1   3   9  11  17   *
*   2   4  10  14  18   *
*   5   6  11  15  19   *
*   7   8  12  16  20   *
*   *   *   *   *   *   *
```

South-East scanning

```
*   *   *   *   *   *   *
*  17  12   9   4   1   *
*  18  13  10   5   2   *
*  19  14  11   6   3   *
*  20  16  15   8   7   *
*   *   *   *   *   *   *
```

South-West scanning

```
*   *   *   *   *   *   *
*  20  16  15  14  13   *
*  19  12  11  10   9   *
*  18   8   7   6   5   *
*  17   4   3   2   1   *
*   *   *   *   *   *   *
```

North-West scanning

```
*   *   *   *   *   *   *
*  16   2   8  14  20   *
*  11  17   3   9  15   *
*   6  12  18   4  10   *
*   1   7  13  19   5   *
*   *   *   *   *   *   *
```

Non-uniform scanning

Figure 1

Note: the symbol * is used to represent boundary points.

THE "GURU" - A MICROPROCESSOR SOLUTION

DR. ORIN E. MARVEL, MR. RAO AYSOLA, AND MR. FRANK ZUHDE
Honeywell
Marine Systems Division
West Covina, California

ABSTRACT

This is the story of a microprocessor development that succeeded because of the flexibility, simplicity, programmability, and ease of implementation of the unit.

The "GURU" was developed to provide control, interface conversion, and stimulation between the DIMUS (AN/BQR-21) Sonar Set and the Development Tester. It was a good 'teacher' of where, when, and how a microprocessor should be used.

INTRODUCTION AND SUMMARY

What do you do when:

1. A complex piece of test equipment has been developed for a product?
2. The test equipment is used to qualify the engineering model?
3. The customer, after successful sea trials of the engineering model, wants a more general-purpose interface?
4. The first production unit rolls off the line and can no longer be connected to the complex test equipment?
5. Documentation is poor to nonexistant?

Enter the programmable microprocessor interface unit which provides flexibility, simplicity, programmability, and ease of implementation.

During the analysis phase, the following requirements constantly pointed to the microprocessor solution. First and foremost, the interface unit must be flexible and accommodate equipment, timing and bit organization changes made after the hardware is built. Secondly, it must be self-testing because of the finger pointing that occurs whenever an error is found. Thirdly, it must verify the test equipments operation without tying up the operational equipment. Fourth, it must perform the data and timing conversions from one interface to the other.

This paper, then, gives an account of the tradeoffs, hardware and software developments, system integration, and lessons learned.

The "GURU" lived up to its name and was a good teacher of where, when, and how a microprocessor should be used.

AN/BQR-21 "DIMUS" SONAR SET

The AN/BQR-21, Figure 1, is a new sonar being developed as part of the Navy Sonar Suite to meet the requirement for broadband passive detection and tracking. It processes and displays data intercepted by a cylindrical, hull-mounted hydrophone array. The AN/BQR-21, with digital signal processing, provides improved broadband performance in both detection and tracking against new and projected threats. Features of the AN/BQR-21 include:

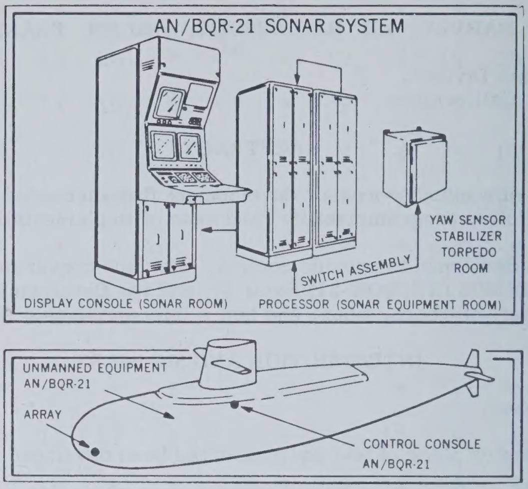

AN/BQR-21 SONAR SYSTEM
YAW SENSOR
STABILIZER
TORPEDO
ROOM
SWITCH ASSEMBLY
DISPLAY CONSOLE (SONAR ROOM)
PROCESSOR (SONAR EQUIPMENT ROOM)
UNMANNED EQUIPMENT
AN/BQR-21
ARRAY
CONTROL CONSOLE
AN/BQR-21

Figure 1. AN/BQR-21

- Increased equipment functions with no increase in personnel
- Preformed beams, providing a higher search rate
- Improved signal processing, allowing detection at lower signal-to-noise ratios
- Higher display data rates, decreasing the time required to detect a target
- Provision of four digital tracking channels, in addition to the retained analog channel, allowing simultaneous tracking of multiple targets of interest
- Increased sonar performance resulting from an improved hydrophone array environment

Simplicity and ease of operation are the characteristics of the AN/BQR-21 system design. Controls and switches have been minimized by the introduction of a keyboard. Displays are arranged for easy visibility. There are three CRT displays: two for presentation of acoustic data, and the third, referred to as the tabloid display, for continuously displaying system status and target data. The tabloid can be shown on any of the three CRT's at the discretion of the operator.

The digital operating mode serves as the primary vehicle for detection. Digital functions are controlled from a small keyboard. Display mode changes, Automatic Tracking activation and deactivation, and other status actions are made by keying the appropriate numeric read from the tabloid. When a numeric has been keyed, the associated function indication on the tabloid blinks to verify the input. The operator may then either execute or cancel his order by pushing the appropriate button.

A separate section of the operator's console is for analog functions. Bearing readout and bearing deviation indication of the analog beam are provided by dials on the control surface and by the tabloid. Analog audio is available to the operator, and a cursor for the analog beam is shown on the video display.

A large number of interface and system growth requirements have lead to the AN/BQR-21 (GPI) General Purpose Interface. This is a semi-handshake interface, where all internal data and status are transmitted every 41 milliseconds. The interface contains 40 lines; of which 4 are control (data clock, data available, ID bit, and user acknowledge) and 36 are data. During an ID word (with the ID bit set) one of the data bits indicates which block of data follows. The timing constraints on the user are that the acknowledge pulse normally must be between 0.5 microseconds and 30 microseconds wide. If the user is acknowledging properly, then the display console will output the next data word within 50 μ seconds. However, if the user fails to acknowledge a data transfer within 50 μ seconds or doesn't acknowledge at all, the GPI goes into a self transfer mode and transmits data whenever it feels like it. After this mode has been entered, recovery can only be made by the user holding acknowledge high for a full 41 millisecond data transfer period.

DIMUS Development Tester (DDT)

The DDT is a piece of test equipment which was designed to provide a means of performing system tests on the AN/BQR-21 sonar. It provides realistic stimuli in the form of target generators and added noise, and monitors the DIMUS system/subsystem operations at numerous points. The DDT also provides for data reduction and recording in the form of hardcopy printout.

The DDT is physically a collection of 6/19-inch wide test racks. It contains stimuli generators, general purpose minicomputer, commercial test equipment, ASR-35 teletype, and terminal printer.

The DDT interfaces consists of three independent autonomous ports. The first is a parallel 40-bit interface containing four control bits (one clock, and a three-bit code defining the data ID), four bits that must always be zero, and 32 data bits. The second is a three-wire serial interface carrying the system status, a data available, and a clock. The third is a 30-bit parallel interface containing six control lines and 24 data lines for track data.

Tradeoffs

It is obvious from the interface descriptions of the last two sections, that the BQR-21/GPI will not talk to the DDT. Actually the solutions to this problem are rather limited because of the customer bias and lack of DDT hardware and software documentation. From the beginning, any modifications to the DDT were prohibited. Thus the only possible solutions are an interface conversion box which is either developed with handwired logic or with a programmable controller.

During the first analysis, the programmable controller had a smaller hardware cost than the logic unit; but the software cost was undefined. Thus, it required a hard sell of management to go with the programmable controller. As it turns out, if we had gone with the handwired logic unit, the whole interface would have been redesigned before the project was over.

Actually the project was typical of aerospace developments where the system definition is happening simultaneously with the hardware design and build. In this case, the actual understanding of the DDT operation only occurred after the GURU (the name given the interface conversion unit) was built and used to stimulate the DDT. Also, because of preproduction pressures no DIMUS unit was available to exercise the DDT, making a handwired logic implementation impossible to use.

Once the programmable controller selection was made, the GURU block diagram, shown later in Figure 2, was developed.

The microprocessor selection can be one of the hardest parts of the whole process, because of the large field of available devices, radical pricing fluctuations, and performance claims and counter claims.

In general the microprocessors can be grouped into the following three categories:

1. Slow and low cost, which use PMOS technology, provide instruction times around 10 microseconds, and cost $5.00 to $20.00 each. These are being sold in four-bit, eight-bit, and bit-slice units.
2. Moderate speed and low cost, which use NMOS technology, provide instruction times around three microseconds, and cost $20.00 to $50.00 each. These are being sold in four-bit and eight-bit units, with 16-bit units available in 1977.
3. High speed and moderate cost, which use bipolar technology (TTL-S mostly), provide instruction times around 200-500 nanoseconds, and cost $100.00 to $300.00 for a chip set that processes 16-bit data words.

The two prominant moderate speed devices are the Motorola 6800 and Intell 8080. Arguments about which is better stretch on into the night and get one nowhere. In my opinion, they are so close to the same that neither is better. My selection of the 6800 for the GURU was then based on the fact that Honeywell had a buying agreement with Motorola for 6800's because of our auto traffic control systems. One should not dwell on the micro selection; but get on to the more serious problems.

The MC6800 is a monolithic eight-bit microprocessor forming the central control function for Motorola's M6800 family. Compatible with TTL, the MC6800, as with all M6800 system parts, requires only one +5.0 volt power supply, and no external TTL devices for bus interface.

The MC6800 is capable of addressing 65k bytes of memory with its 16-bit address lines. The eight-bit data bus is bidirectional, as well as three-state, making direct memory addressing and multiprocessing applications realizable.

- Eight-Bit Parallel Processing
- Bidirectional Data Bus
- Sixteen-Bit Address Bus--65k Bytes of Addressing
- 72 Instructions--Variable Length
- Seven Addressing Modes--Direct, Relative, Immediate, Indexed, Extended, Implied, and Accumulator
- Variable Length Stack
- Vectored Restart
- Maskable Interrupt Vector
- Separate Non-Maskable Interrupt--Internal Registers Saved In Stack
- Six Internal Registers--Two Accumulators, Index Register, Program Counter, Stack Pointer, and Condition Code Register
- Direct Memory Addressing (DMA) and Multiple Processor Capability
- Clock Rates as High as 1 mHz
- Simple Bus Interface Without TTL
- Halt and Single Instruction Execution Capability

The basic EXORciser (M68SDT) consists of the MPU Module, the Debug Module, the Baud Rate Module, the Power Supply and the chassis. These modules are built around the M6800 Microcomputer Family of Parts (MC6800 Microprocessing Unit, MCM6810 Random Access Memory, MC6820 Peripheral Interface Adapter, MCM6830 Read Only Memory, and MC6850 Asynchronous Communication Interface Adapter devices).

The MPU Module incorporates the MC6800 Microprocessing Unit (MPU) and the system clock. This module serves a dual function in that it provides the MPU and clock for both the EXORciser Debug and the user's system. With the exception of the EXORciser clock, the system restart, and the DBE signal (delayed), the MPU Module appears exactly like the MC6800 Microprocessing Unit with unlimited TTL bus drive capability. The MPU is an eight-bit parallel device capable of addressing 65,536 bytes of memory. In addition, the MPU addresses its input and output devices as memory. The MPU also provides the EXORciser with 72 variable length instructions and the capability of responding to real time interrupt signals.

The Debug Module provides the EXORciser with the capability to evaluate and debug the user's program. The three MCM6830 ROM memories on this module contain the EXbug Firmware that provides the EXORciser with its unique software control features. The module also has two MCM6810 RAM memories to provide a scratch pad memory to the EXbug firmware.

The Debug Module working with the Baud Rate Module provides the EXORciser with eight standard baud rates between 110 and 9600 Baud.

These modules also interface the EXORciser with a TTY or RS232C compatible terminal and provide a reader control signal for remote control of modified manual TTY terminals.

The Power Supply provides the EXORciser with the +5 Vdc, +12 Vdc, and -12 Vdc power sources to support a full EXORciser rack of modules.

The chassis is capable of holding 14 plug-in modules. The Power Supply and Baud Rate Modules are not plug-in modules but mount directly to the chassis. Two versions of the chassis are available--the rack mounted version and the table top version.

The references used for these developments are:

1. "M6800 Microprocessor Applications Manual"
2. "M6800 Microprocessor Programming Manual"
3. "M6800 EXORciser User's Guide"

The GURU development used the following seven steps:

1. Develop a word statement of the problem (specification).
2. Draw a flow diagram for the design (algorithm).
3. Design the hardware interfaces as required.
4. Define the subroutines required and memory map requirements.
5. Write the subroutines and main program.
6. Load the total program and debug the system.
7. Update, complete, and release the documentation.

GURU HARDWARE

Figure 2 is a detailed block diagram of the GURU. The hardware is designed around the Motorola 6800 chip set and was limited to the parts contained in a M6800B kit. This includes:

- 1 - 6800 MPU
- 2 - 6810 RAM
- 2 - 6820 PIA

The biggest mistake in this development was the limitation of two PIA's. The PIA is actually the most versatile I/O device that I have ever seen or used. It contains two (eight-bit) input/output ports which are defined as inputs or outputs at program execution time through a control word. The control word also determines the polarity of the handshake and transition edge that sets the input flag or interrupt. The PIA can also be used as a temporary storage register because the output data is register "buffered" and can be recalled into the MPU. This feature makes a "wrap around" test very easy. The test data can be sent out one PIA port (which also stores the data in the PIA register) and brought into the A register through the other PIA port. Then the A register can be compared with the PIA register.

The operation of the GURU revolves around the two PIA's. PIA-I is used for data input and output from the 6800. PIA-I's flag is set every time the GPI clock occurs and is reset when the 6800 MPU tests the status. PIA-I's output clock is used for gating data into the selected output register. PIA-II is used for all discrete type controls. These include:

- Normal/Test Switch Setting
- Test 1/Test 2 Switch Setting
- Page/Test Switch Setting
- STA Input ID
- ITA Input ID
- BA Input ID
- Track Input ID
- Serial Output Data Clock
- Parallel Output Data Clock
- Three lines that the decoded to select the appropriate input mux position (36 bits are brought into the MPU as 5 - 8 bit words) or the acknowledged response.
- Four lines that are decoded to select the appropriate output register or to light one of the lights.

I will never again decode or mux PIA outputs. PIA's are too inexpensive and easy to use to pay the complexity, confusion, added software, inflexibility, and timing price of decoding/muxing the PIA outputs.

Since the 6800 has only one data/address bus, the PIA's are connected and treated like memory. Thus, a memory map shows the RAM's at the lower end of memory, the PIA's in the middle, and the program memory PROM's in the upper part of memory.

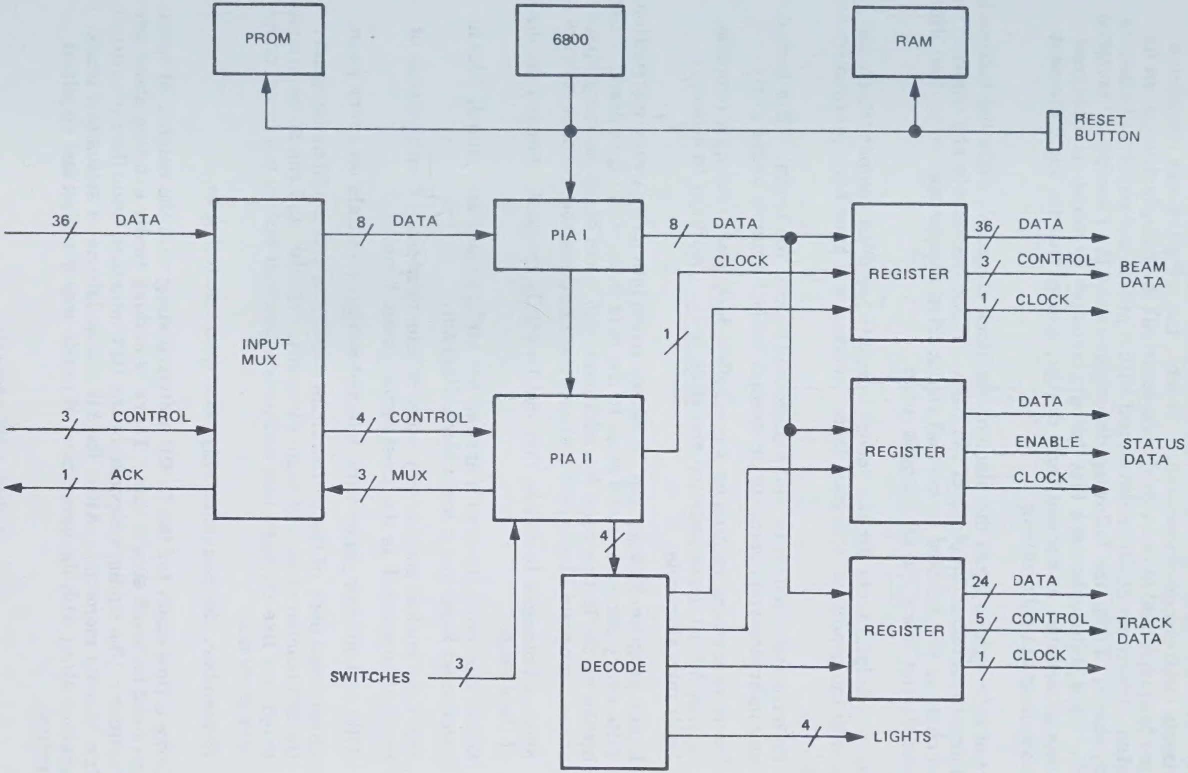

PROM
6800
RAM
RESET BUTTON
36 DATA
3 CONTROL
1 ACK
INPUT MUX
8 DATA
4 CONTROL
3 MUX
PIA I
PIA II
8 DATA
CLOCK
1
4
SWITCHES
3
DECODE
REGISTER
36 DATA
3 CONTROL
1 CLOCK
BEAM DATA
REGISTER
DATA
ENABLE
CLOCK
STATUS DATA
REGISTER
24 DATA
5 CONTROL
1 CLOCK
TRACK DATA
4 LIGHTS

Figure 2. GURU Block Diagram

GURU SOFTWARE

A different tact than normal was taken on the software development because it was being done by hardware people. First of all, I do not believe in special software aids for a simple control microprocessor. In fact, the software was completed using machine language before one could even learn all the complexities of an assembler. Also, programs could be developed with a pencil and paper at home, in a car, or anywhere. Thus, the following description uses the concept of "programmable logic". This philosophy says that "programmable" systems are designed using the same procedures as normal logic design, except that the logic element blocks are now program subroutines.

A flow chart of subroutines takes the place of the block diagram, while the subroutine listing of instructions takes the place of the LBD. These documents are engineering released and changes are handled by normal engineering change methods. Red lining should be used during debug for all changes made.

The following are helpful hints for the "programmable logic" developer which are synonomous with the practices of a good logic designer and have been learned from experience:

1. Be sure the problem is well understood before you begin. This includes any interface requirements or special speed/memory usage goals.
2. Treat everything written as a controlled document. Set up a common format for all documentation and stick to it. Don't get in a hurry. Don't take shortcuts.
3. In any program listing, put functional descriptions by every instruction. Title every subroutine and stick to the title every time it is used. Leave blanks on the listing page for additions; and leave blank memory space around program blocks/subroutines for additions and to correct errors.
4. Make all changes in red so they can be easily tracked. Record the date of the change.
5. Show connecting lines with arrows for all branches and jumps, like interconnect lines and a logic block diagram.
6. Don't get rushed and be very wary of interruptions. Continuations of work after a break is the most error-prone time.
7. Title and number pages with date and designer initials on every page.
8. Understand each of the instructions used (like you would understand the intricacies of an MSI logic element). "Build" and run some simple programs like memory test and instruction test before you begin complex systems.
9. Remember, the computer only does what you tell it to.

Figure 3 shows a flow chart for the GURU software along with the number of memory locations used for each subroutine. There is nothing really exciting about any of the subroutines. The timing response of the GPI interface described previously gave us a few anxious moments. Also, the bit manipulations of status and track word conversions along with the generation of parity was complex and required careful attention.

GURU OPERATION

The GURU latched to the 6800 Exorciser for program and interface development is shown in Figure 4. The GURU operates with its own self-contained PROM memory. For stimulating and debugging the DDT and operational verification of the software,

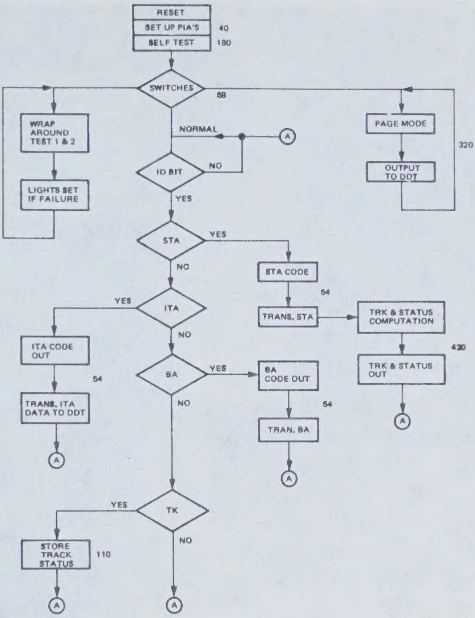

RESET
SET UP PIA'S
40
SELF TEST
180
SWITCHES
68
WRAP AROUND TEST 1 & 2
LIGHTS SET IF FAILURE
40
NORMAL
A
ID BIT
NO
YES
PAGE MODE
OUTPUT TO DDT
320
STA
YES
NO
STA CODE
54
TRANS. STA
TRK & STATUS COMPUTATION
430
TRK & STATUS OUT
A
ITA
YES
NO
ITA CODE OUT
54
TRANS. ITA DATA TO DDT
A
BA
YES
NO
BA CODE OUT
54
TRAN. BA
A
TK
YES
NO
STORE TRACK STATUS
110
A
A

Figure 3. Software Flow Chart

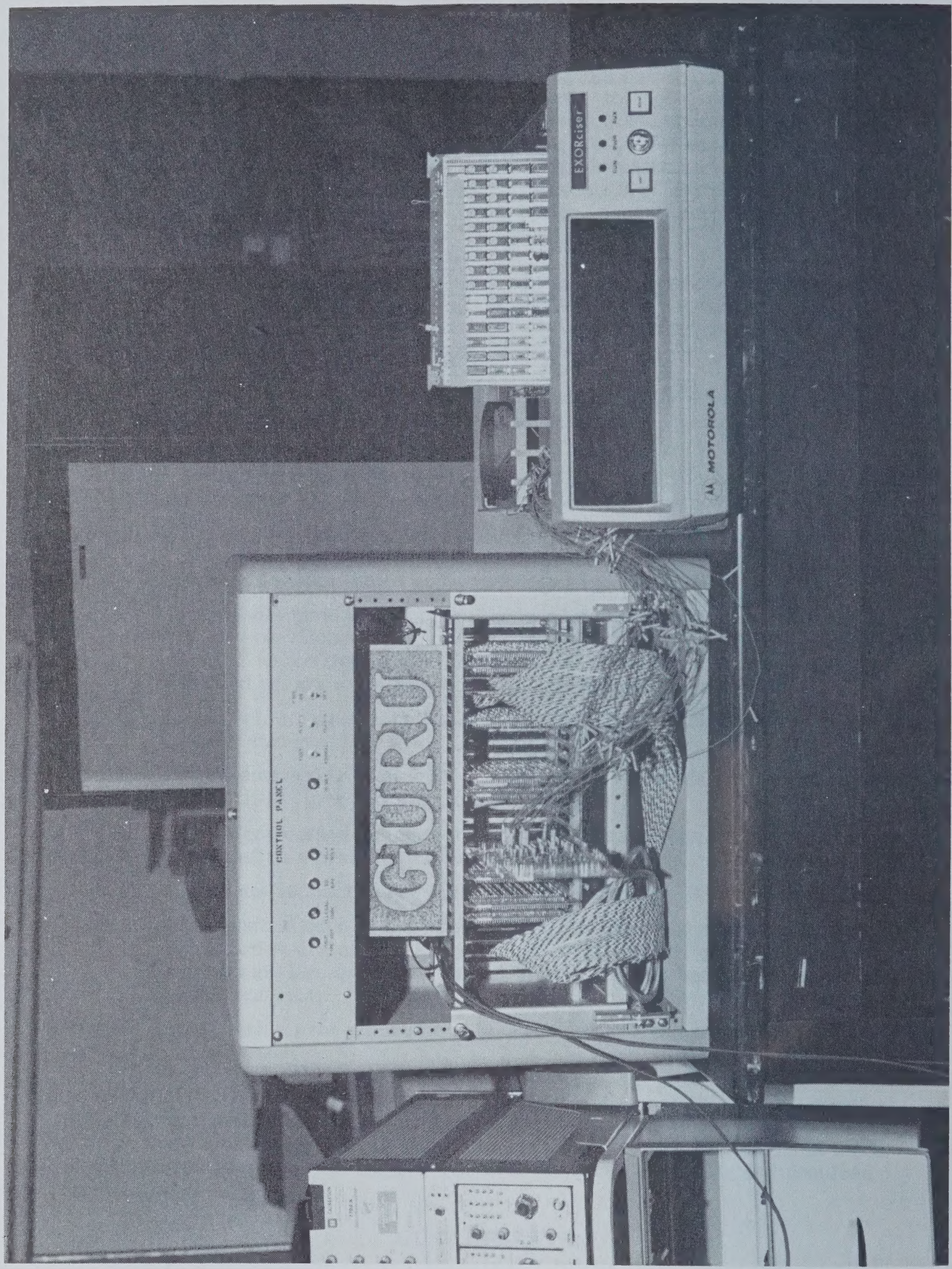

GURU
CONTROL PANEL
EXORciser
MOTOROLA

Figure 4. GURU With Exorciser

the Exorciser was latched to the GURU. This was done by paralleling the lines from the 6800 MPU socket in the GURU to the blank socket of the 6800 in the Exorciser. Thus, the 6800 MPU was actually in both systems. By interfacing with the RAM and PIA's, it operated the GURU. By interfacing with the Exorciser bus, it used the 10K Exorciser memory for program storage and the ex-bug card for TTY interfacing, software debug, and operational debug. This configuration was used for a month while the DDT was debugged and its operation understood. It was a good thing that PROMS had not been initially blown because the software was completely revised after this initial period. I am definitely convinced that the total flexibility of a programmable interface controller should be kept as long into a program as possible (even up to delivery).

The GURU has three operating modes. Mode changeover is made by toggle switch manipulation and pushing the reset button - Figures 3 and 4. The four lights on the control panel have the following meaning:

- The "Program Time-Out" light is operated by a one shot set to about one second. If the program even hangs up and does not retrigger the one shot, the light will come on.
- The "Illegal Code" light is used to indicate that an illegal ID or status condition has been found.
- The "No GPI" light stays on until the clock from the GPI interface is recognized and then it is turned off.
- The "Self-Test" light will come on if a self-test error is found. The self-test program is performed every time the reset button is pushed.

The first mode, Wrap-Around Test, is entered by placing the switches in test/page off/test 1 - test 2. By placing a cable from the output beam data port to the input port (Figure 2), a wrap-around self test is performed. By placing a cable from the output status/track data ports to the input port, a second wrap-around self test is performed. These two tests insure that all the GURU circuitry is operating correctly. These subroutines were also used in the initial checkout of the GURU box.

The second mode, Page, is entered by placing the switches in test/page on. In this mode, predetermined data blocks are continuously sent to the DDT for each of the output data types. When the DDT prints out this data, we are confident of the GURU and DDT operation. This subroutine was also used to stimulate the DDT during its debug phase.

The third mode, Normal Operation, is entered by placing the first switch in normal. This mode is best understood by looking at the flow chart of Figure 3. Whenever the ID bit is encountered, the ID word is searched for Short Term Average (STA), Intermediate Term Average (ITA), Batch Average (BA), and Track/Status (T/S) data. All other data types from the BQR-21 are ignored. When ITA and BA data blocks are recognized, the ID code is added and the data transferred directly to the DDT. When the T/S data block is recognized, the data is stored in the GURU RAM memory. After the STA data blocks have been transferred to the DDT, the track/status computations are performed on the previously stored data. The new data blocks are then transmitted to the DDT. The track/status computations are performed after STA because there is always a five millisecond gap following STA while data is transmitted from the BQR-21 that is not used by the DDT. Actually, the track/status words are transmitted every other cycle because the DDT cannot accept them as fast as every cycle. This last minute modification was easy in the software, but would have been very difficult in a fixed wired logic unit. The programmability and versatility of the GURU has been proven in cost and schedule savings many times during this development.

TECHNIQUES FOR REDUCING PROCEDURAL DEPENDENCY IN A PROGRAM

LAWRENCE S. CHEUNG
Department of Electrical Engineering
Marquette University, Milwaukee, WI

FREDERIC J. MOWLE
Department of Electrical Engineering
Purdue University, W. Lafayette, IN

ABSTRACT

In a program the amount of parallelism available is largely limited by the logical dependencies among its instructions. A set of program simplification techniques is designed to deal with this problem. The techniques discussed are evaluated by means of a simulator with the results showing significant improvement.

Introduction

In a computer system with pipelined processor and/or multiple functional units (e.g. CDC7600), the amount of parallelism available in a program is limited by the dependencies among its instructions [7]. In general, the dependencies can be classified into 3 groups, procedural dependency caused by jump instructions, operational dependency caused by a lack of resources, and data dependency caused by the logical sequencing of a calculation.

Various approaches [4,6,8,9,10] have been used to reduce the dependency problems. Data dependency is inherited from the algorithm used, and can only be resolved by introducing new algorithms for parallel machines. Operational dependency can be reduced by providing more hardware. Procedural dependency can be minimized by simplifying the program structures. This paper addresses itself to the problem of procedural dependency by introducing a set of transformations which will reduce the number of jump instructions executed.

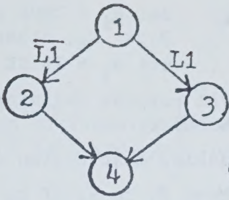
Elimination of Jump Instructions Using Dynamic Algebra

Procedural dependency is a major cause of inefficiency in a parallel processing system [11]. In analyzing this problem, it is convenient to view a program as a control flow graph [1], i.e., a labelled, directed graph with the basic blocks of the program identified with nodes and control paths among these blocks identified with directed edges. A basic block is a linear sequence of program instructions having one entry point (the first instruction executed) and one exit point (the last instruction executed). All programs may be analyzed as a sequence of basic blocks.

The significant contribution which dynamic algebra [3,9] has brought to parallel processing is a systematic method for reducing the number of jump instructions in a program. An example demonstrating this concept is shown in Fig. 1.

It is assumed in Fig. 1 that L1 either has a value of all 1's or all 0's depending on whether it represents a true or false value. Conditions are associated with some nodes of the graph. They are used to indicate that if the condition is false, the corresponding node is equivalent to a no operation. Conditions are also associated with some of the edges of the graph. They indicate the correct path to be taken when there is more than one alternative.

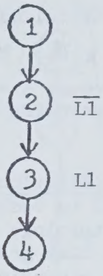
```
    IF (A.GT.B.) GO TO 1
    X=X+Y/Z
    GO TO 2
  1 X=Y
    Y=Z
  2 . . . .
```



(a) Original program

```
    L1=A.GT.B
    X=(L1.A.X ).OR.(.N.L1.A. (X+Y/Z))
    X=(L1.A.Y) .OR.(.N.L1.A.X)
    Y=(L1.A.Z) .OR.(.N.L1.A.Y)
```

(b) Program obtained after conversion



```
    L1=A.GT.B
    X=(L1.A.Y ).OR.(.N.L1.A.(X+Y/Z))
    Y=(L1.A.Z ).OR.(.N.L1.A.Y)
```

(c) Program obtained after simplification

Fig. 1. Example illustrating the concept of dynamic algebra

In general, the advantages obtained from using dynamic algebra are:

1) It is simple and straight forward;
2) Jump instructions can be avoided, and
3) Better code may be obtained when several nodes are merged together to form a single node.

The gain obtained from using dynamic algebra is not free. In Fig. 1, the expression X+Y/Z has to be evaluated even when A is less than B. Assuming they are equally probable, the time and resources spent evaluating that expression is wasted half of the time. A net loss in time may result if calculations are time consuming.

To determine whether a transformation is worthwhile or not is difficult, since it is not known exactly how much time is gained by eliminating a conditional jump or how much time is lost by doing some unnecessary operations. This is entirely machine and program dependent. If a system has an unlimited supply of resources, the cost to do some unnecessary operations is negligible; but if the resources are limited, the extra operations performed may actually slow down the program. Intuitively speaking, the smaller the number of operations in the nodes involved, the greater is the gain, since this puts less burden on the system. For simplicity, we choose not to carry out a transformation if the number of operations involved exceeds a predetermined number. This scheme can be modified by assigning different weights to different operations with the decision based on the weighted sum.

Although dynamic algebra succeeds in removing "go to" statements from a program, some of the statements introduced may contain logical variables and, when expand into machine codes, may require jump instructions in many machines (consider the statement L = I.GT.J). New machine instructions must be added to deal with logical variables. Some examples are

SGE R_i R_j R_k Set R_i = TRUE (11··11)
if $R_j \geq R_k$; otherwise
set R_i = FALSE (00··00)

SL R_i R_j R_k Set R_i = TRUE if $R_j < R_k$; otherwise set R_i = FALSE

R_i may be a general purpose register or a register designated for logical variables only.

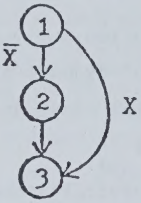
Another useful machine instruction is the conditional move instruction

MV_m R_i R_j R_k Move R_i to R_j if R_m = TRUE, otherwise move R_i to R_k.

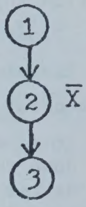
The statement X = (L.A.Y). OR.(.N.L.A.Z) can be expanded into MV1 R2 R3 R4 assuming L, X, Y, Z are stored in R1, R2, R3 and R4 respectively.

Program Simplification Techniques

Figures 2 to 7 show a series of transformations which can be used to simplify the topology of many programs thereby reducing the number of jump instructions. The transformations are illustrated using directed graphs with examples given in Fortran.

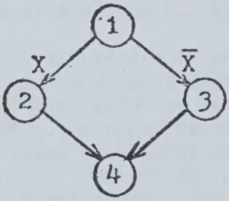


```
   IF(A.GT.B) GO TO 1
   C=D
 1 . . . .
```

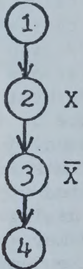


```
X=A.GT.B
C=C.A.X.OR.D.A..N.X
```

Fig. 2. Transformation 1

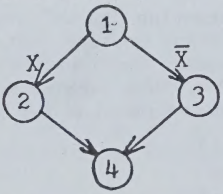


```
   IF(A.GT.B) GO TO 2
   C=D
   GO TO 4
 2 D=E
 4 . . . .
```

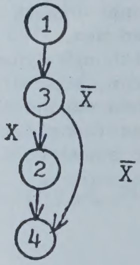


```
X=A.GT.B.
C=C.A.X.OR.D.A..N.X
D=E.A.X.OR.D.A..N.X
```

Fig. 3. Transformation 2



```
   IF (A.GT.B) GO TO 2
   C=D
```



```
X=A.GT.B
C=C.A.X.OR.D.A..N.X
```

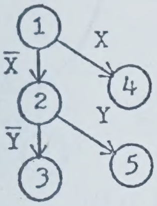
Fig. 4. Transformation 3

```
      GO TO 4
    2 . . . .
      . . . .
    4 . . . .
```

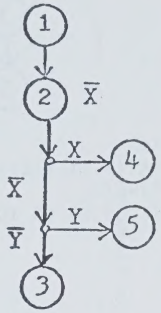
```
      IF(.N.X) GO TO 4
      . . .
      . . .
    4 . . .
```

Fig. 4. Transformation 3 (continued)

Transformation 3 is a variation of Transformation 2 and is useful when the number of instructions in node 3 is large, but that in node 2 is small.


1
X
$\overline{X}$
4
2
Y
$\overline{Y}$
5
3

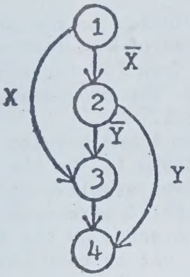
```
      IF(A.GT.B) GO TO 4
      C=D
      IF(E.GT.F) GO TO 5
```


1
2
$\overline{X}$
X
4
$\overline{X}$
Y
5
$\overline{Y}$
3

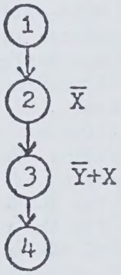
```
      X=A.GT.B
      C=C.A.X.OR.D.A..N.X
      Y=E.GT.F
      IF(X)  GO TO 4
      IF(Y)  GO TO 5
```

Fig. 5. Transformation 4

Transformation 4 delays the execution of the conditional jump in node 1 until node 2 is executed. The objective is to group all the conditional jumps together so that there will be more time to evaluate the conditions before the jumps take place. If information is available on how often the conditions are true or false, the order of testing the conditions may be rearranged to give a better performance (see Fig. 7). Sometimes, it is possible to eliminate a conditional jump whenever node 4 is the same as node 5 or node 3 (see Fig. 6).

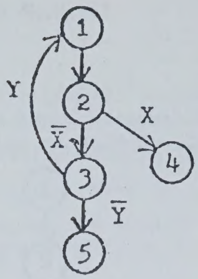

1
$\overline{X}$
X
2
$\overline{Y}$
Y
3
4

```
      IF(A.GT.B) GO TO 3
      C=D
      IF(E.GT.F) GO TO 4
    3 E=F
    4 . . . .
```

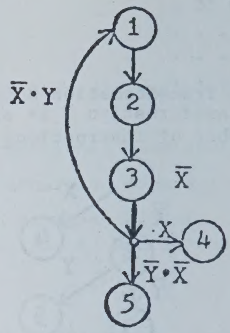

1
2
$\overline{X}$
3
$\overline{Y}$+X
4

```
      X=A.GT.B
      C=C.A.X.OR.D.A..N.X
      Y=X.OR.(E.LE.F)
      E=(F.A.Y).OR.(E.A..N.Y)
```

Fig. 6. Transformation 5



```
1 I=I+1
  IF(A(I).GT.B(I)) GO TO 4
  A(I)=B(I)
  IF(I.LT.N) GO TO 1
```



```
1 I=I+1
  X=A(I).GT.B(I)
  A(I)=A(I).A.X.OR.B(I).A.
                       .N.X
  IF((I.LT.N).A.(.N.X)) GO
                        TO 1
  IF(X) GO TO 4
```

Fig. 7.Transformation 6

Transformation 5 removes the conditional jumps from node 1 and node 2.

Transformation 6 removes a conditional jump out of a loop and is an important and productive Transformation. Notice that the order of execution of the two conditional branches are reversed since it is assumed that the probability of jumping out of a loop is small.

The Experiment

In order to evaluate the suggested transformations, a set of programs was selected and compiled using two different approaches. Version 1 was compiled using all the transformations suggested in this paper, while version 2 did not use any. Since an automatic compiler was not available, the compilation was done by hand. In order to minimize any bias, both versions were optimized as much as possible.

A computer with multiple functional units similar to the structure of CDC 7600 was designed and simulated. It consisted of one instruction fetch unit and 8 distinct classes of functional units (see [5] for a detail description). No vector operations were allowed. The programs were run on the simulator and the results are shown in Table 1.

Comparing the execution times of the version 1 and version 2 programs, an improvement of 93% was observed as shown in Table 1. On the average, 48.8% of the jumps are eliminated using the various simplification techniques while the penalty paid is the execution of 17% more code. Note that for VSORTM, 55% more code is executed in order to speed up the program by 11%. This shows, in some cases, the pay off may not be very large. The memory requirements for the first version of the sample program are slightly larger than those of the second version. On the average, there is a difference of 12%.

One measure on the effective utilization of a memory is the memory space execution time product (MTP). Table 1 also compares the MTP's of the two versions of the programs. On the average, the first version of the samples uses the memory 84% more effective than the second version.

Table 1. Characteristics of the sample programs

		IE_i	IE_1/IE_2	J_i	J_1/J_2	M_i	M_1/M_2	MTP_2/MTP_1	T_i	T_2/T_1
AL1	Ver. 1	271	1.19	19	.50	65	1.14	1.45	172	1.66
	Ver. 2	227		38		57			285	
VSORTM	Ver. 1	850	1.55	101	.81	123	1.21	.918	968	1.11
	Ver. 2	548		127		102			1075	
DCSIQU	Ver. 1	2587	.926	127	.56	157	1.09	2.18	1258	2.37
	Ver. 2	2842		226		144			2985	
ZBRENT	Ver. 1	1479	1.07	86	.38	121	.945	2.61	1249	2.46
	Ver. 2	1234		227		128			3075	
VMULFS	Ver. 1	2925	1.02	258	.46	39	1.05	1.72	1294	1.81
	Ver. 2	2871		564		37			2339	
BSORT	Ver. 1	1710	1.29	117	.36	46	1.31	1.79	1166	2.14
	Ver. 2	1321		321		35			2493	
	Mean		1.17		.51		1.12	1.84		1.93

IE_i – Instructions executed in version i of the programs

J_i – Jump instructions executed in version i of the programs

M_i – Core requirement for version i of the programs

MTP_i – Memory space - execution time product for version i of the programs

T_i – Execution time for version i of the programs (in cycles)

Summary

A set of compiler optimization techniques for computers with multiple functional unit was proposed. Their main objective is to increase the amount of parallelism in a program by simplifying the structure and by rearranging the ordering of an instruction sequence. Their performance was evaluated using a simulator and significant improvement was obtained.

REFERENCES

1. Allen, Francis E., "Control flow analysis", SIGPLAN Notices - 5, 7 (July 1970), Proceedings of a Symposium on Compiler Optimization, 1-20.

2. Bernstein, A. J., "Analysis of programs for parallel processing", IEEE Trans. Elect. Computer, EC-15, 5 (Oct. 1966), 757-763.

3. Boyce, Raymond F., "Topological reorganization as an aid to program simplification", Ph.D. thesis, Dept. of Computer Science, Purdue University, 1972.

4. Chamberlin, Donald D., "The single-assignment approach to parallel processing", 1971 Fall Joint Computer Conf., AFIPS Cong. Proc., Vol. 39, AFIPS Press, Montvale, N.J.: AFIPS Press, 263-269.

5. Cheung, L. S., "Techniques for reducing dependencies among instructions for a parallel single processor computer system", Ph.D. thesis, Dept. of Electrical Engineering, Purdue University, 1975.

6. Clark, E. R., "On the automatic simplification of source language programs", Comm. ACM, 10, 3 (March 1967), 160-165.

7. Flynn, M. J., "Some computer organization and their effectiveness", IEEE Trans. Computer, C-21, 9, (Sept. 1972), 948-960.

8. Nievergelt, J., "On the automatic simplification of source language programs", Comm. ACM, 8, 6 (June 1965), 366-370.

9. Noonan, Robert E., "Computer programming with a dynamic algebra", Ph.D. thesis, Department of Computer Science, Purdue University, 1971.

10. Riseman, E.M. and Foster, C. C., "The inhibition of potential parallelism by conditional jumps", IEEE Trans. Computer, C-21, 12 (Dec. 1972), 1405-1411

MULTI - VALUED ASYNCHRONOUS SEQUENTIAL MODULES

ANTHONY S. WOJCIK
Department of Computer Science
Illinois Institute of Technology
Chicago, Illinois 60616

ABSTRACT

The design of multi-valued speed-independent asynchronous sequential modules is considered. The basic state assignment scheme is presented along with the necessary control primitives. The general monitoring and control structure that allows the interconnection of complex modules-both combinational and sequential - is described.

INTRODUCTION

A topic of continuing interest in the field of switching theory is the design of asynchronous networks. In particular, the design of speed-independent asynchronous circuits has been investigated by many researchers, including [1-7]. The use of multi-valued signals in such circuits has been alluded to and attempted in the past [2,3,9,10]. However only the successful work reported in [6-8,11-12] appears to be based on a realistic model of non-binary asynchronous behavior and to be of general applicability.

The purpose of this paper is to describe some extensions of the results initially presented in [6], which are reported in detail in [7], to asynchronous modules that incorporate sequential circuits.

BASIC PRINCIPLES

Multi-valued asynchronous operation is based on a special interpretation of the three-valued Post algebra in which the logic values are labelled and linearly ordered as $0<1<2$. The extreme elements, 0 and 2, are treated as binary information values while the intermediate element, 1, is used as a control value. Any change from information level 0(2) to information level 2(0) must first pass through the control level. This behavioral requirement allows for the design of hazard-free circuits and the use of technologically feasible logic elements [7,8,11,12].

Table 1 describes the basic logic operators that are needed. These operators are used to construct a multi-valued JK flip-flop and the three basic control primitives, described in Figure 1. A key aspect of the design of combinational networks using these operators and the control primitives is that, essentially, the logic designer uses standard binary design procedures to develop the non-binary networks. (It should be noted that the logic values 0 and 2 together with the operators MAX,MIN, IDENTITY and STRONG-NEGATION forms a 2-valued Boolean algebra.) Although multi-valued logic elements and control primitives and three distinct logic levels are used, the designer need only understand the functional behavior of the elements and primitives and the intended control functions of the primitives in order to incorporate them in the design of a combinational module. The multi-valued elements and primitives need not be designed. Figure 2 illustrates the structure of a combinational logic module. In Figure 2, the Data Processing Network is designed using binary design procedures with respect to the 0 and 2 information levels; the control level 1 does not enter the design process.

SEQUENTIAL CIRCUITS

Let us now consider the design of modules that contain a data processing network that is a sequential circuit. Although research has been reported on the design of non-binary asynchronous sequential circuits [8,12], the techniques and procedures are truly "multi-valued" and require the designer to master new, rather complex, design procedures. The approach to be briefly described here uses a method that allows the designer to incorporate binary, asynchronous sequential logic design techniques.

First, the binary, restricted single transition time (restricted STT) state assignment is used [4,13]. This asignment allows several secondary variables to change simultaneously during a state transition without causing critical races. Further, both the primary and secondary outputs of the combinational logic part of the circuit are uniquely determined by the primary input variables and only those secondary variables which remain fixed during a state transition. This feature allows certain hazards to be avoided and provides a means for properly monitoring the status of the circuit.

Since this binary state assignment is being used, the basic three-valued JK flip-flop is placed into a new control primitive, the FLIP-FLOP CONTROL PRIMITIVE, described in Figure 3. This is necessary in order to provide for the necessary monitoring and control of the data processing network. Referring to Figure 3, J^E and K^E are the actual excitation functions that will result from using the binary state assignment and the logic design rules - described in the next paragraph. The unary operators that provide inputs to J^A and K^A of the flip-flop itself mask out any steady state logic 1 values that may appear on J^E and K^E. This is necessary since the assumed operational behavior of a network of logic modules provides that each module that is processing data (all primary inputs are 0s or 2s) is preceeded and succeeded by modules that are processing the control signal (all inputs are 1). Hence, a sequential module needs to remember a previous data value while it currently processes the control signal. Figure 1 shows that steady state 1 signals as inputs to the flip-flop would force it into the 1 state, thus destroying all record of the previous data value. Using the FLIP-FLOP PRIMITIVE allows the entire circuit to process and respond to three levels of signal while taking advantage of the binary state assignment and binary design procedures, Hence, the FLIP-FLOP primitive is used whenever a memory element for the sequential circuit is required.

To design the sequential circuit, only the values 0 and 2 attained by the primary and secondary inputs need to be considered. The following design rules are followed to implement the output functions of the sequential circuit:

1. In forming an excitation table for the J_i and K_i excitation values,
    a. if in column j and row r of the transition table $y_i = 0(2)$ and $Y_i = 0(2)$, set $J_iK_i = 02(20)$;
    b. if in column i and row r of the transition table $y_i = 0(2)$ and $Y_i = 2(0)$, set $J_iK_i = 20(02)$.
2. In forming the algebraic equations for the J_i and K_i excitation functions and the primary output functions,
    a. if I_i and I_j are two adjacent primary input vectors, no term in any of the equations can be formed that assumes the value 2 - in a sum-of-products form for the functions (0 - in a product-of sums form for the functions) - for both of the inputs I_i and I_j;
    b. the design must ensure that each equation attains the value 1 at the control signal.

These rules are necessary to ensure that certain hazards do not occur and that the circuit can be properly monitored. Further, logic circuits that incorporate these rules can be shown to function properly under the delay assumptions associated with speed-independent design [7].

Figure 4 illustrates the structure of a module incorporating a sequential circuit. The internal monitoring of the circuit is only slightly different from the monitoring required for a combinational circuit, Figure 2. The key point is that modules communicate via the C_p and C_s status lines and by the primary inputs, the X_is. This uniform interconnection feature allows both sequential and combinational modules to be incorporated in a network.

CONCLUSIONS

The results that have been briefly reported in this paper indicate that multi-valued signals can be incorporated in the design of asynchronous, speed-independent modules. Further, the designer may use binary design procedures with respect to the binary information levels of 0 and 2 without explicitly taking into account the control 1 level. By using four basic control primitives, the required monitoring and control of a network composed of the interconnection of modules is obtained. Further study is needed to determine if a simpler control structure can be obtained and to investigate the design of multi-valued modules under different interpretations of the meanings of the available logic levels.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation under Grant DCR 74-01409.

REFERENCES

[1] D. E. Muller, "Asynchronous logics and application to information processing," Proc. Symp. on Application of Swit. Theory in Space Tech., Aiken and Main, Eds., Stanford University Press, Stanford, Calif. pp. 289-297, 1963.

[2] D. Hammel, "Ideas on asynchronous feedback networks,"Proc. 5th Ann. Symp. on Swit. Circuit Theory and Logical Design, Princeton University, pp. 4-11, November 1964.

[3] B. Elspas, et al., "Investigation of propagation-limited computer networks," Final Reports - Phases 1, 11, 111, Stanford Research Institute, April 1964, July 1965, June 1966.

[4] D. B. Armstrong, A. D. Friedman, and P. R. Menon, "Design of asynchronous circuits assuming unbounded gate delays, "IEEE Trans. Comp., Vol. C-18, pp. 1110-1120, December 1969.

[5] R. M. Keller, "Towards a theory of universal speed-independent modules," IEEE Trans. Comp., Vol. C-23, pp 21-33, January 1974.

[6] A. S. Wojcik, "Multi-valued asynchronous modules," Proc. 13th Ann. Allerton Conf. on Circuit and System Theory, October 1975.

[7] A. S. Wojcik, "The design of multi-valued asynchronous modules," Report of the Department of Computer Science, Illinois Institute of Technology, Chicago, to appear.

[8] S. J. Sheafor, "The design of multiple-valued asynchronous sequential circuits," Coordinated Science Laboratory Report R-663, University of Illinois at Urbana-Champaign, Sept. 1974.

[9] A. S. Wojcik, "An application of ternary logic to logic sequencing," Coordinated Science Laboratory Report R-388, University of Illinois, Urbana, July 1968.

[10] H. Mine, et al., "Self-timing asynchronous logical networks using 3-valued NAND circuits and its effect on a binary adder," Report of the Faculty of Engineering, Kyoto University, Kyoto-shi, Japan, 1968.

[11] A. S. Wojcik, "Multi-valued asynchronous circuits," 1973 Int. Symp. on Multiple-Valued Logic, University of Toronto, May 1973.

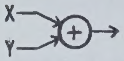
[12] A. S. Wojcik, "Multi-valued asynchronous sequential circuits," 1974 Int. Symp. on Multiple-Valued Logic, West Virginia University, May 1974.

[13] S. H. Unger, Asynchronous Sequential Switching Circuits, Wiley, New York, 1969.

Table 1. Basic logic operators.

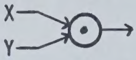
+	Y		
X	0	1	2
0	0	1	2
1	1	1	2
2	2	2	2

MAX(X,Y)



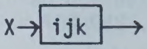
•	Y		
X	0	1	2
0	0	0	0
1	0	1	1
2	0	1	2

MIN(X,Y)



X	$X^{(ijk)}$
0	k
1	j
2	i

UNARY OPERATORS



X	$X^{(210)}$
0	0
1	1
2	2

IDENTITY OPERATOR

X	$X^{(012)} = \overline{X}$
0	2
1	1
2	0

STRONG-NEGATION OPERATOR

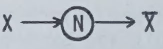
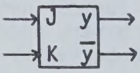


Figure 1. JK flip-flop and basic control primitives.

JK Flip-Flop

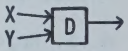
y	JK 00	01	02	10	11	12	20	21	22
0	0	0	0	1	1	1	2	2	2
1	1	1	0	1	1	1	2	1	1
2	2	1	0	2	1	0	2	1	0

$$Y = J \cdot \overline{K} + \overline{K} \cdot y + J \cdot \overline{y}$$



D Primitive

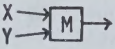
	Y		
X	0	1	2
0	0	1	0
1	2	1	1
2	0	1	0



$$D(X,Y) = X^{(020)} \cdot Y^{(020)} + 1 \cdot (X^{(020)} + Y^{(020)})$$

M Primitive

	Y		
X	0	1	2
0	0	1	1
1	1	1	1
2	1	1	2



$$M(X,Y) = X \cdot Y + 1 \cdot (X + Y)$$

CONTROLLER Primitive

$$J = 1 \cdot S^{(200)} \cdot f \cdot C_s^{(002)} + y^{(220)} \cdot \overline{S} \cdot f \cdot C_s$$

$$K = 1 \cdot S^{(200)} \cdot \overline{f} \cdot C_s^{(002)} + y^{(022)} \cdot \overline{S} \cdot \overline{f} \cdot C_s$$

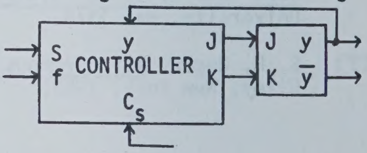


Figure 2. Structure of a combinational logic module.

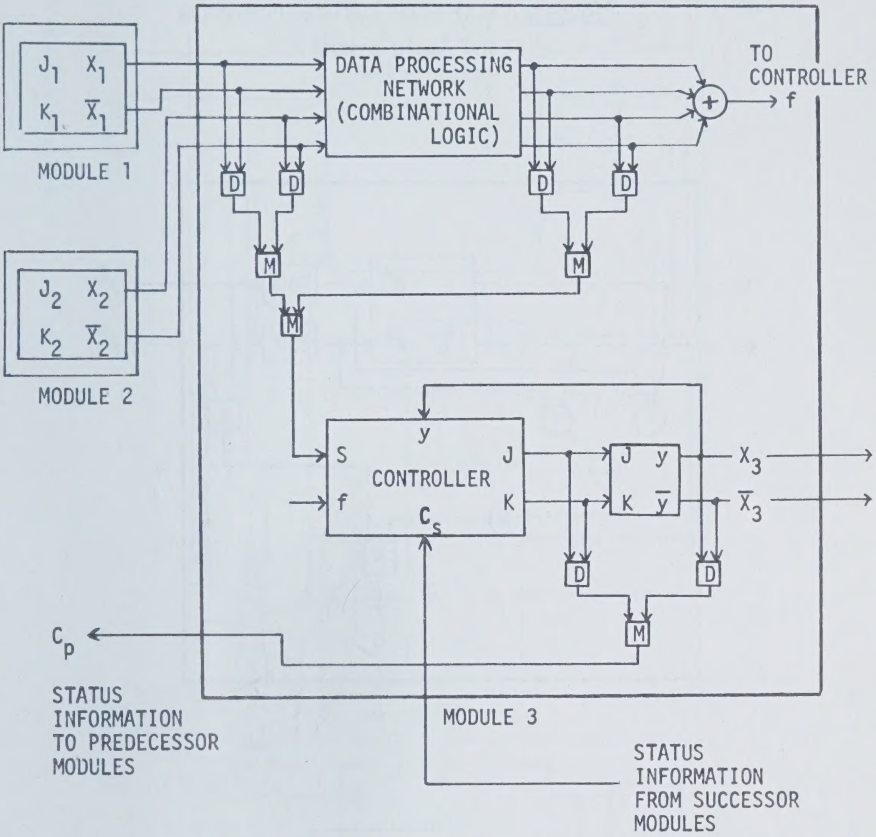
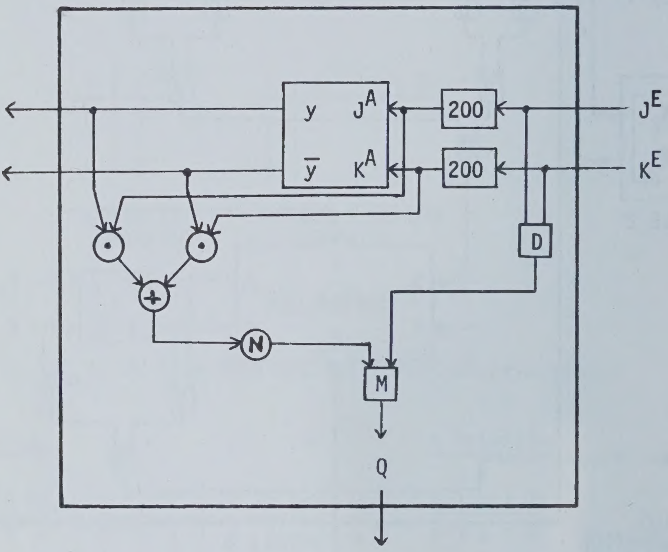

J_1 X_1
K_1 $\overline{X}_1$
MODULE 1
J_2 X_2
K_2 $\overline{X}_2$
MODULE 2
DATA PROCESSING NETWORK (COMBINATIONAL LOGIC)
TO CONTROLLER f
+
D
M
y
S
CONTROLLER
J
K
f
C_s
J y
K $\overline{y}$
X_3
$\overline{X}_3$
C_p
STATUS INFORMATION TO PREDECESSOR MODULES
MODULE 3
STATUS INFORMATION FROM SUCCESSOR MODULES

Figure 3. FLIP-FLOP CONTROL Primitive and logic symbol.



y
J^A
ȳ
K^A
200
200
J^E
K^E
D
N
M
Q

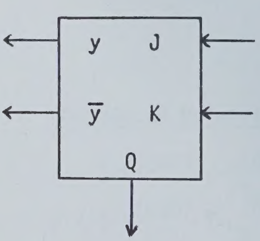
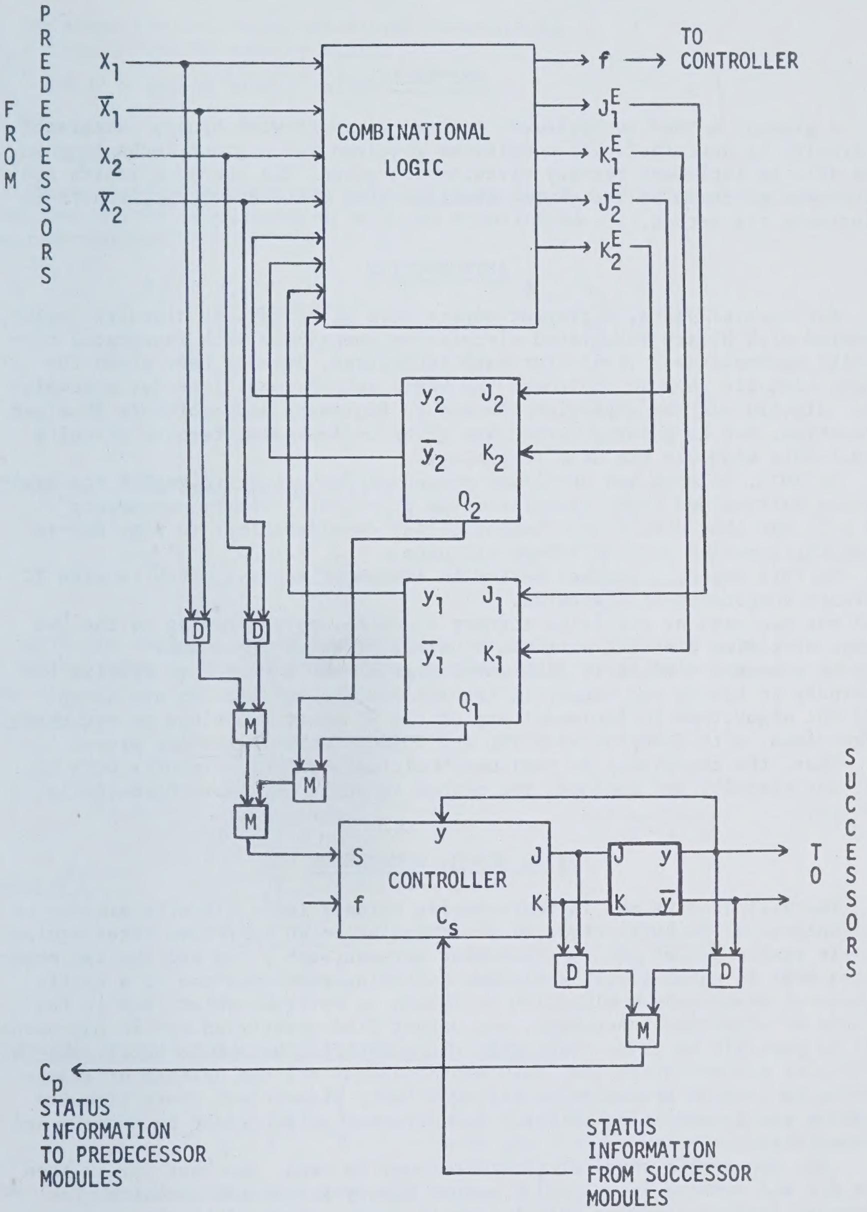


Figure 4. Structure of a sequential logic module.


FROM PREDECESSORS
X_1
$\overline{X}_1$
X_2
$\overline{X}_2$
COMBINATIONAL LOGIC
f
TO CONTROLLER
J_1^E
K_1^E
J_2^E
K_2^E
y_2
J_2
$\overline{y}_2$
K_2
Q_2
y_1
J_1
$\overline{y}_1$
K_1
Q_1
D
D
M
M
M
CONTROLLER
S
f
y
J
K
C_s
J
y
K
$\overline{y}$
TO SUCCESSORS
D
D
M
C_p
STATUS INFORMATION TO PREDECESSOR MODULES
STATUS INFORMATION FROM SUCCESSOR MODULES

A GENERAL METHOD TO IMPLEMENT TERNARY CIRCUITS WITH BINARY COMPONENTS

Daniel ETIEMBLE
UNIVERSITE PARIS 6 - FRANCE

Michel ISRAEL
CNAM PARIS - FRANCE

ABSTRACT

A general method to implement ternary circuits with binary integrated circuits is described. The conditions required for a given technology to be able to implement ternary circuits are given. The needed circuits and the general formulas are given. Examples with TTL and COSMOS circuits illustrate the method.

INTRODUCTION

For several years, different papers have given ternary circuits implemented with binary integrated circuits or compatible with integrated circuits technologies. (1-4). For each technology, authors have given the gate circuits able to implement the basic set of operations for a complete algebra and the expansion theorem to implement any arbitrary 3-valued function. But no general method was given to implement ternary circuits avalaible whatever the used technology.

In 1974, JE Birk and DE Farmer presented "An algebraic method for designing multivalued logic circuits using principally binary components". (5). But they didn't give technological considerations to show how to use their method with IC binary circuits.

In this paper, a general method to implement ternary circuits with IC binary components is presented.

1) the two ways of realizing ternary circuits, corresponding to the two ways of coding 3 states with two 2 valued circuits are shown.

2) by a method similar to Birk and Farmer's one, a method to realize the ternary to binary and binary to ternary conversion circuits are shown.

3) the algorithms to implement any of the 27 unary functions or any n-ary functions, with examples with TTL and COSMOS technologies are given.

Thus, the conditions to implement multivalued logic circuits with IC binary circuits are defined. The method to build the basic circuits is given.

USING BINARY COMPONENTS

The difficulties met in implementing ternary logic circuits are due to technological factors. Physical phenomena selected must have three equiprobable stable states and the distances between each state and the two other ones must be equal to get a minimum switching time. The use of a cyclic physical phenomenom would allow to obtain an optimum system. But in the field of electrical phenomena, one cannot find quantified cyclic phenomena. It is possible to prove that using non quantified phenomena would make the circuits unserviceable (6). They would present all the defects of linear circuits : noise transmission and additivity properties. These circuits adding and transmissing internal and external noises would be thereby non iteratives.

So, non cyclic physical phenomena must be used. The best way is then to use and combine in a special manner binary integrated circuits. The ternary logical circuits will be apt to answer to the following requirements : rapidity, simplicity of circuits, minimum costs, integrability...

We shall first examine the different ways of coding ternary representations with binary circuits.

TERNARY CODED BINARY REPRESENTATIONS.

We adopt the following notational conventions.

$E_3 = \{0,1,2\}$ Set of ternary values.

$E_2 = \{0,1\}$ Set of binary values

$z \in E_3$

$x,y \in E_2$

There are $C_4^3 = 4$ different manners to choise 3 different states among the four states corresponding to x and y variables.

Representations :

R_1

x_1	0	0	1
y_1	0	1	1

R_2

x_2	0	1	1
y_2	0	0	1

R_3

x_3	0	0	1
y_3	0	1	0

R_4

x_4	0	1	1
y_4	1	0	1

Let be x^* the dual variable of x
y^* y

R_1^*

x_1^*	1	1	0
y_1^*	1	0	0

$R_1 \equiv R_2$

R_3^*

x_3^*	1	1	0
y_3^*	1	0	1

$R_3 \equiv R_4$

With the dual representations, there are only 2 different ways to choise 3 states among 4. The minimum codes corresponding to these representations are :

Code n°1

x	0	0	1
y	0	1	1
z	0	1	2

or

x	0	1	1
y	0	0	1
z	2	1	0

Code n°2

x	0	0	1
y	1	0	0
z	0	1	2

or

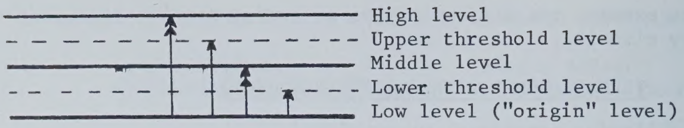
x	1	1	0
y	0	1	1
z	0	1	2

Physical representation corresponding to code n°1

x	0	0 ▌	1
y	0 ▌	1	1
z	0	1	2

The state corresponding to z = 0 (0,0) may be considered as the origin state. There are 2 different threshold levels for x and y.

The corresponding physical model will be :


High level
Upper threshold level
Middle level
Lower threshold level
Low level ("origin" level)

The dual physical model is also available.
These two physical models are called the "non symetrical" ones.
Example : the TTL ternary inverter (3) belongs to this model.

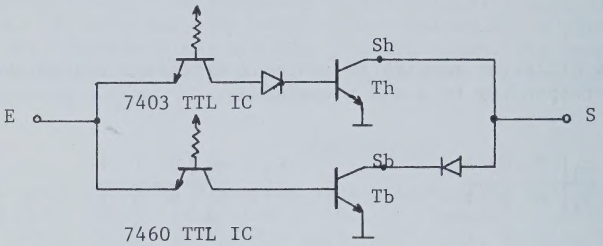

Sh
Th
7403 TTL IC
E
S
Sb
Tb
7460 TTL IC

fig.1

The threshold levels for the transistors T_b ans T_h are respectively V_D and V_{2D}

V_D corresponds to one base emittor junction threshold level.
V_{2D} corresponds to two base emittor junctions threshold level.

E	T_b	T_h
$E < V_D$	off	off
$V_D < E < V_{2D}$	on	off
$E > V_{2D}$	on	on

Let be $e \in E_3$ and $x,y \in E_2$

$e = 0 \leftrightarrow E < V_D$
$e = 1 \leftrightarrow V_D < E < V_{2D}$
$e = 2 \leftrightarrow E > V_{2D}$

T_h off $\leftrightarrow x = 0$
T_h on $\leftrightarrow x = 1$
T_b off $\leftrightarrow y = 0$
T_b on $\leftrightarrow y = 1$

e	x	y
0	0	0
1	0	1
2	1	1

To the 3 valued imput e corresponds two 2 valued variables x and y which are implemented with the TTL 7403 and 7460 IC. The 3 valued output is got by a combinational circuit : the output diode.

e	x	y	T_h	T_b	S
0	0	0	off	off	High level
1	0	1	off	on	Middle level
2	1	1	on	on	Low level

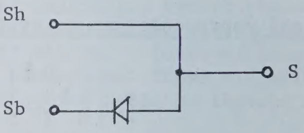


fig 2

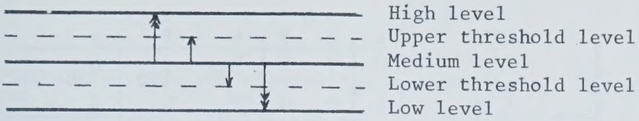
Physical representation corresponding to code n°2

x	0	0	1
y	1	0	0
z	0	1	2

The state z = 1 (0,0) may be considered as the origin state. The two different threshold for x and y are symetrical with reference to the state (0,0)

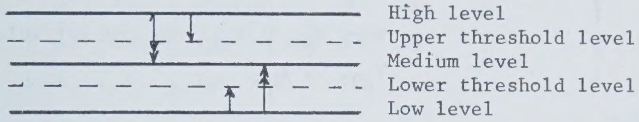
The two corresponding physical models are :

a)


High level
Upper threshold level
Medium level
Lower threshold level
Low level

the medium level is the reference one.

b)


High level
Upper threshold level
Medium level
Lower threshold level
Low level

the high and low levels are the two reference Ones.

These two physical models are called the "symetrical" ones.
Example : the COSMOS ternary inverter (4) belongs to this model.

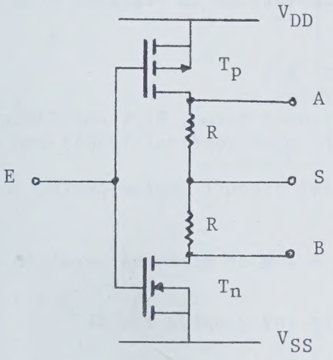

V_{DD}
T_p
A
R
E
S
R
B
T_n
V_{SS}

fig. 3

Let V_p and V_n the threshold levels for the p and n channel transistors.

$$V_{DD} - E < V_p \leftrightarrow T_p \quad \text{off}$$
$$V_{DD} - E > V_p \leftrightarrow T_p \quad \text{on}$$
$$E - V_{SS} > V_n \leftrightarrow T_n \quad \text{on}$$
$$E - V_{SS} < V_n \leftrightarrow T_n \quad \text{off}$$

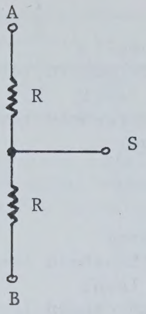
$$E < V_{SS} + V_n \leftrightarrow e = 0$$
$$V_{SS} + V_n < E < V_{DD} - V_p \leftrightarrow e = 1$$
$$E > V_{DD} - V_p \leftrightarrow e = 2$$

T_n on $\leftrightarrow$ $x = 0$
T_n off $\leftrightarrow$ $x = 1$
T_p on $\leftrightarrow$ $y = 0$
T_p off $\leftrightarrow$ $y = 1$

e	x	y
0	0	1
1	0	0
2	1	0

Two 2 valued variables x and y correspond to the 3 valued input e. They are implemented with the p and n channel transistors of COSMOS IC.
Combinational 3 valued output circuits.

$$S = \frac{A + B}{2}$$



e	x	y	A	B	S
0	0	1	V_{DD}	V_{DD}	V_{DD}
1	0	0	V_{DD}	V_{SS}	$\frac{V_{DD} + V_{SS}}{2}$
2	1	0	V_{SS}	V_{SS}	V_{SS}

When $V_{SS} = -V_{DD}$, the out put values are V_{DD}, $0, -V_{DD}$

fig. 4

IMPLEMENTATION OF CONVERSION CIRCUITS

The used methode is similar to the methods presented by Birk and Farmer (5,6). The ternary to binary conversion is realized with threshold detectors.

$$x^i = \begin{cases} 1 & \text{if } x \geqslant i \\ 0 & \text{otherwise} \end{cases}$$

For the sum (maximum) of product terms, Birk and Farmer use level restoring binary AND gates (at the next to final level) and finally a max gate at the output.

For the sum, we use a special circuit corresponding to the used technology.

$f = 2.f_2+1.f_1+0.f_0$ with f_i = sum of minterms equal to i.

The function f is determined by any 2 uples fj, fk

f	f_2	f_1	f_0
0	0	0	1
1	0	1	Ø
2	1	Ø	Ø

Ø is the notational term for don't care. With these don't cares, the usual covering algorithms may be used to synthetise the f function.

a) "non symetrical" representation : example with TTL technology.
We use two 7460 (or 7403) IC in the following way.

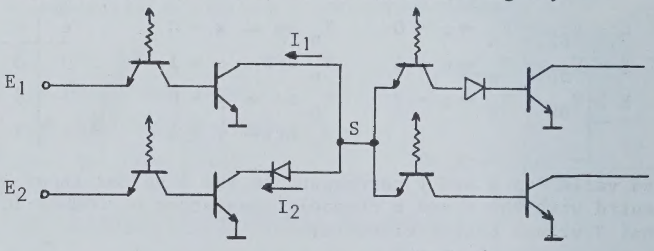


fig.5

$$E_i < V_D \leftrightarrow I_i = 0 \qquad e_i = 0$$
$$E_i > V_D \leftrightarrow I_i \neq 0 \qquad e_i \neq 0$$

We may determine the relations between e_1, e_2, and f_2, f_1, f_0.

e_1	e_2	I_1	I_2	S	s	f_2	f_1	f_0
0	0	=0	=0	High	2	1	Ø	Ø
0	1	=0	≠0	Midlle	1	0	1	Ø
1	0	≠0	=0	Low	0	0	0	1
1	1	≠	≠	Low	0	0	0	1

For the 3 2-uplets f_j, f_k, the relations are :

2-uplet f_2, f_1

e_1	e_2	f_2	f_1
0	0	1	Ø
0	1	0	1
1	Ø	0	0

$e_1 = \overline{f}_1 . \overline{f}_2$
$e_2 = \overline{f}_2$

2-uplet f_2, f_0

e_1	e_2	f_2	f_0
0	0	1	Ø
0	1	0	0
1	Ø	0	1

$e_1 = \overline{f}_2 . f_0$
$e_2 = \overline{f}_2$

2-uplet f_1, f_0

e_1	e_2	f_1	f_0
0	0	0	0
0	1	1	Ø
1	Ø	0	1

$e_1 = \overline{f}_1 . f_0$
$e_2 = f_1$

We may choise any of the 3 2-uplets to get the final "sum circuit".
With f_2 f_1, the sum circuit is given fig 6

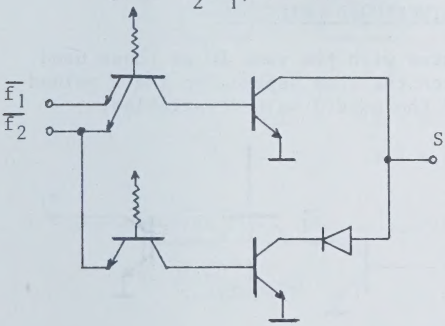


fig.6

when $\overline{f}_1 . \overline{f}_2 = \overline{f}_1$
$(\overline{f}_1 \leqslant \overline{f}_2)$

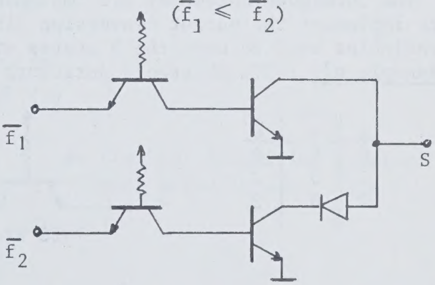
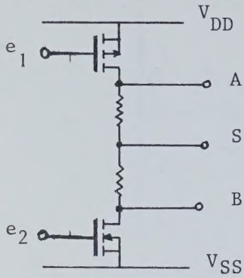


fig.7

b) "Symetrical" representation : example with COSMOS technology


V_{DD}
e_1
A
S
B
e_2
V_{SS}

e_1	e_2	x	y	A	B	S	s
0	0	0	1	V_{DD}	V_{DD}	High	2
0	1	0	0	V_{DD}	V_{SS}	Mid	1
1	0	1	1	?	?	?	?
1	1	1	0	V_{SS}	V_{SS}	Low	0

fig.8

e_1	e_2	s	f_2	f_1	f_0
0	0	2	1	0	0
0	1	1	0	1	0
1	1	0	0	0	1

The formulas are the same as the TTL ones.

$e_1 = \bar{f}_2 . \bar{f}_1$ $\quad$ $e_1 = \bar{f}_2 . f_0$ $\quad$ $e_1 = \bar{f}_1 . f_0$

$e_2 = \bar{f}_2$ $\quad$ $e_2 = \bar{f}_2$ $\quad$ $e_2 = f_1$

With f_2 f_1, the "sum circuit" is when $\bar{f}_2 \geqslant \bar{f}_1$, then

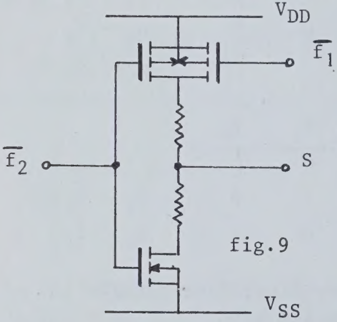

V_{DD}
$\bar{f}_1$
$\bar{f}_2$
S
V_{SS}

fig.9

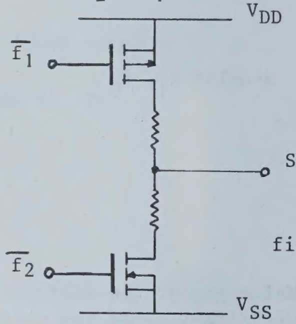

V_{DD}
$\bar{f}_1$
S
$\bar{f}_2$
V_{SS}

fig.10

IMPLEMENTATION OF THRESHOLD DETECTORS

The threshold detectors are implemented with the same IC as those used to implement the output conversion circuits. They depend on the 2 valued variables used to code the 3 states of the used 3 valued variables.

Example n°1 : TTL threshold detectors

e	$\bar{e}^1$	e^1
0	2	0
1	0	2
2	0	2

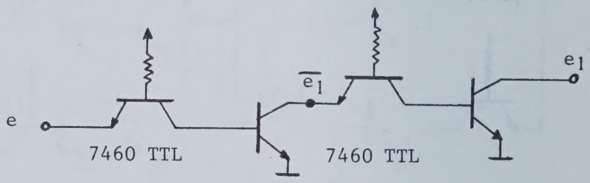

e
7460 TTL
$\bar{e}_1$
7460 TTL
e_1

fig.11

e	$\overline{e^2}$	e^2
0	2	0
1	2	0
2	0	2

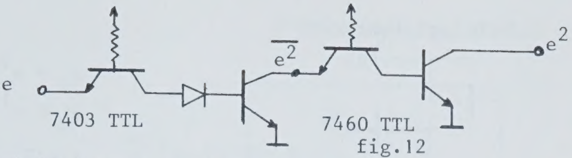

e
7403 TTL
$\overline{e^2}$
7460 TTL
e^2

fig.12

Example n°2 : COSMOS threshold detectors

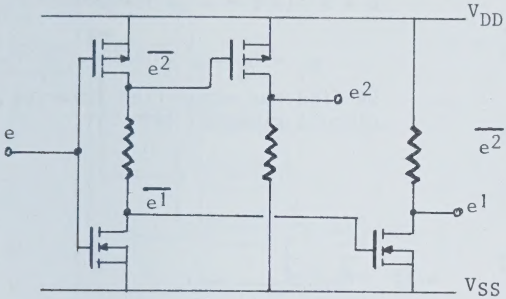

V_{DD}
e
$\overline{e^2}$
e2
$\overline{e1}$
e1
V_{SS}

fig. 13

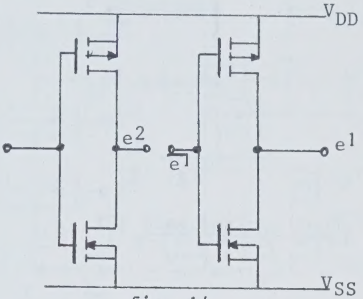

V_{DD}
$\overline{e2}$
e2
$\overline{e1}$
e1
V_{SS}

fig. 14

IMPLEMENTATION OF TERNARY FUNCTIONS

The implementation of any ternary function is possible with any technology where the threshold detectors and the "sum" circuits are implemented. The synthesis of f_2, f_1, or f_0 functions are realized with the well known algorithms for 2 valued operators. The simplifications corresponding to a technology must be determined specifically.
We shall give the examples of implementing two unary functions (inverter and Post cycling gate) and one binary function (comparator).

Inverter :$f_{2\ 1\ 0}$

$f_2 = x^1$

$f_1 = x^2$ with $\overline{f}_1 \leqslant \overline{f}_2$

TTL implementation

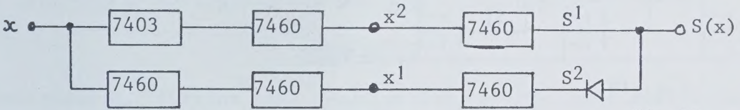
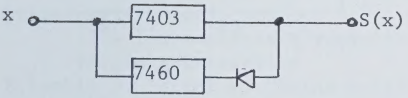

x
7403
7460
x^2
7460
S^1
S(x)
7460
7460
x^1
7460
S^2

fig.15

$$S_1\ (x) = \overline{x^1}\ (x^2)$$

$$\text{but } x^{\overline{1}}\ (x^2) \equiv x^{\overline{1}}\ (x)$$

$$S_2\ (x) = x^{\overline{1}}\ (x^1) \equiv x^{\overline{1}}\ (x)$$


x
7403
S(x)
7460

We find the simplified inverter circuit given in (3)

fig. 16

COSMOS implementation :

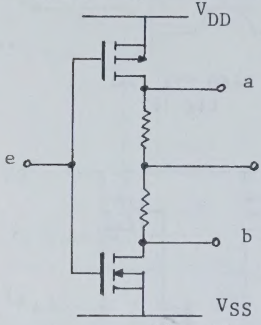

V_{DD}
a
e
S
b
V_{SS}

fig. 17

$$e_1 = x^2$$
$$e_2 = x^1$$
$$a = \bar{x}^2\,(e^1) = \bar{x}^2(x^2) = \bar{x}^2(x)$$
$$b = \bar{x}^1\,(e^2) = \bar{x}^1(x^1) = \bar{x}^1(x)$$

$$e_1 = e_2 = x$$

We find the simplified inverter circuit given in (4)

Post cycling gate : $f_{1\,2\,0}$

$f_2 = x^1 . \bar{x}^2$ $\quad \bar{f}_2 = \overline{x^1 . \bar{x}^2}$ $\quad$ with $\bar{f}_1 \leqslant \bar{f}_2$

$\bar{f}_1 = x^2$

TTL implementation

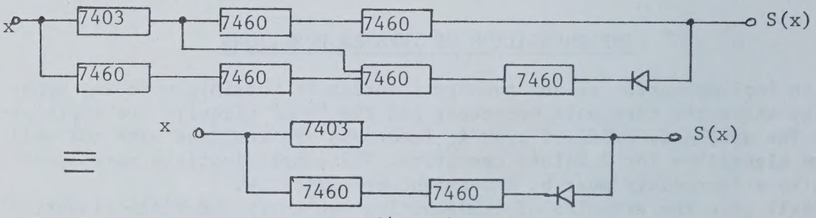
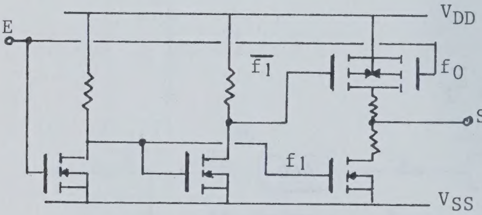

x
7403
7460
7460
7460
7460
7460
7460
S(x)
≡
x
7403
S(x)
7460
7460

fig. 18

COSMOS implementation


E
V_{DD}
$\bar{f}_1$
f_0
S
f_1
V_{SS}

$$f_0 = x^2$$
$$f_1 = \bar{x}^1$$
$$\bar{f}_1 = x^1$$

fig. 19

Comparator

y \ x	0	1	2
0	1	2	2
1	0	1	2
2	0	0	1

$$f_2 = \bar{y}^1 . x^1 + x^2 . \bar{y}^2$$
$$\bar{f}_2 = \overline{\bar{y}^1 . x^1} \;.\; \overline{x^2 . \bar{y}^2}$$
$$f_1 = \bar{y}^1 + x^2 + x^1 . \bar{y}^2$$
$$\bar{f}_1 = y^1 . \bar{x}^2 . \overline{x^1 . \bar{y}^2}$$

TTL implementation

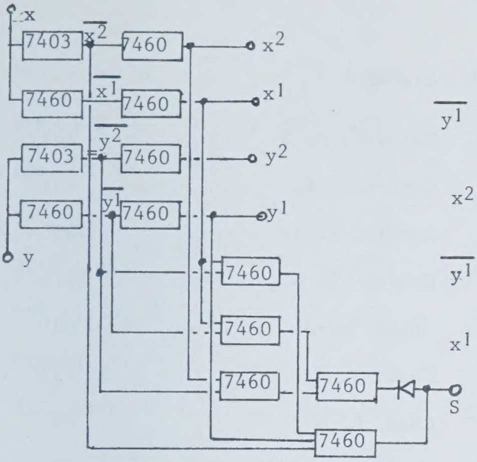

x
7403
$\overline{x^2}$
7460
x^2
7460
$\overline{x^1}$
7460
x^1
7403
$\overline{y^2}$
7460
y^2
7460
$\overline{y^1}$
7460
y^1
y
7460
7460
7460
7460
7460
S

fig. 20

COSMOS implementation

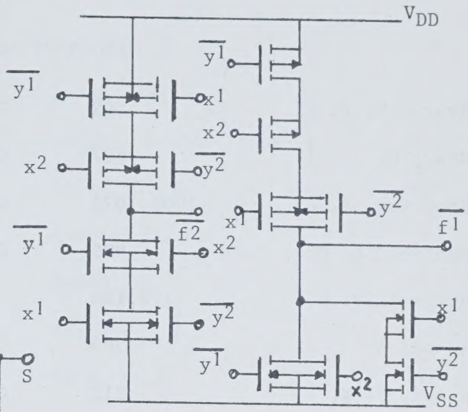

V_{DD}
$\overline{y^1}$
x^1
$\overline{y^1}$
x^2
x^2
y^2
x^1
$\overline{y^2}$
$\overline{f^1}$
$\overline{f2}$
$\overline{y^1}$
x^2
x^1
$\overline{y^2}$
x^1
$\overline{y^1}$
$x2$
$\overline{y^2}$
V_{SS}

fig. 21

CONCLUSION

The paper has defined the conditions required to implement ternary circuits with binary components. These components must be able to realize two different threshold levels for "non symetrical" circuits, or two symetrical threshold level with "symetrical" circuits. Then "combinational" circuits must be found to get three stable states. It is easy when these two requirements are found to realize the threshold detector circuits and the output binary to ternary conversion circuit. Our method allows to implement any 3 valued function as soon as the threshold detectors and the output conversion circuit are defined for a given technology. The general method used with simplification algorithms leads to minimum circuits. For instance, the Cosmos cycling gate (fig. 19) has less channel p and n Mos transistors as Mouftah's cycling gate (4). This work will be completed with further investigations with other technology (IIL...) and descriptions of the simplification expressions for each technology.

REFERENCES

(1) Dunderdale H. "Current mode circuits for ternary logic realizations"
Electronics Letters Vol 5 n° 23 pp575 577
(2) Dunderdale H. "Current Mode Circuits for the unary functions of a ternary variable".
Electronics Letters Vol 6 n° 1 pp 15 16
(3) Etiemble D Israel M " A new concept for ternary logic elements"
(4) Mouftah H. Jordan IB "Integrated circuits for ternary logic".
(5) Birk JE Farmer IB "Design of multivalued combinational circuits using principally binary components "
1974 International Symposium on Multiple Valued Logic. West Virginia University
(6) Etiemble D. Israel M. "Noise margins in multivalued logic circuits"
1976 ISMVL. Utah University.

INDEX OF AUTHORS

Abbrecht, P. H. 509
Acar, C. 587
Adedotun, E. A. 1007,1023
Agarwal, L. C. 905
Agarwal, V. K. 178,188
Ahmad, A. 569
Allen, J. W. 211
Alvarado, F. L. 279
Ardema, M. D. 807
Asner, B., Jr. 328
Aysola, R. 1113
Barmish, B. R. 644
Basso, R. J. 652
Berman, L. E. 1084
Bertsekas, D. 636
Beutler, F. J. 19
Bhargava, D. K. 915
Bhargava, V. K. 731
Bhat, K. P. M. 550
Bien, Z. 662
Binding, P. 819
Blankenship, C. L. 832
Bock, M. E. 349
Boroson, D. M. 517
Brown, H. E. 299
Calise, A. J. 793
Ćalović, M. S. 389
Cambanis, S. 451
Carlin, H. J. 1033
Cheung, L. S. 1124
Chien, M.-J. 881
Chow, J. H. 132
Chua, L. O. 667
Chyung, D. H. 662
Cobb, J. D. 1042
Cohn-Sfetcu, S. 418
Conan, J. 695
Costello, D. J., Jr. 694
Daaboul, M. M. 578
Datko, R. 62
Davison, E. J. 339,482,983,1052
Dayal, G. 269
DeClaris, N. 753
DeSantis, R. M. 964
Delfour, M. 535
Đorović, M. V. 389
Dunning, L. A. 704
Dussault, J. 189
Eastwood, L. F., Jr. 781
Ellis, J. B. 905,915
Enns, M. 289

Etiemble, D. 1140
Eyman, E. D. 1007, 1023
Fang, T.-P. 27
Fam, A. T. 380
Fogel, A. 944
Fraley, D. W. 626
Fruend, E. 113
Gallagher, N. C., Jr. 211, 460
Gejji, R. R. 1091
Gesing, W. E. 1052
Giuli, D. 732
Gnanasekaran, R. 243
Goldenberg, A. 983
Gutmann, R. L. 147
Haccoun, D. 695
Haddad, A. H. 841
Hakimi, S. L. 712
Halanay, A. 328
Han, Y.-W. 616
Hazony, D. 677
Helton, J. W. 91
Hemmati, F. 694
Heydt, G. T. 253
Hibey, J. L. 470
Ho, Y. C. 338
Huang, S. T. 451
Ionescu, T. 122
Israel, M. 1140
Jacobs, M. Q. 528
Javid, S. H. 825
Kadota, T. T. 10
Kamen, E. W. 310
Kanefsky, M. 740
Kang, S. M. 667
Kashyap, R. L. 491
Kassam, S. A. 954
Kazakos, D. 787
Keskar, D. A. 1063
Ketelsen, M. 156
Khalil, H. K. 851
Khan, K. A. 569
Kinney, L. L. 616
Kokotović, P. V. 841
Koralek, R. W. 558
Kuhl, J. G. 191
Kumar, P. R. 781
Kumar, R. 896
Kuszta, B. 1016
Kwong, R. H. 545
Leake, R. J. 652, 792
Lee, E. B. 308
Lefebvre, S. 964

Leshem, G. 222
Li, R. 763
Liu, R. 423
Ma, R. K. T. 408
Manitius, A. 320
Marvel, O. E. 1113
Masson, G. M. 178, 188
Mayeda, W. 772
Mayer, D. J. 233
Medanić, J. 369
Meditch, J. S. 380
Melsa, J. L. 500
Mese, E. D. 732
Metze, G. 189
Metzner, J. J. 722
Meyer, G. G. L. 1102
Michel, A. N. 70, 147
Miller, R. K. 70
Mintz, M. 349
Mitra, S. K. 243
Molnar, C. E. 27
Morefield, C. L. 1074
Morley, R. E., Jr. 963
Mowle, F. J. 1124
Muaki, H. 683
Mudge, T. 606
Murata, T. 597
Newcomb, R. W. 753
Ntafos, S. C. 712
Onaga, K. 772
Papanicolaou, G. C. 832
Papantoni-Kazakos, P. 763
Patel, J. N. 481
Patel, R. 974
Peach, L. C. 481
Penm, H.-W. 740
Pickel, W. C. 528
Poor, H. V. 432, 925
Powers, W. F. 509
Radojevic, D. 369
Rajbman, N. S. 998
Reddy, S. M. 191
Reese, A. C. W. 558
Reynolds, D. A. 166
Rhodes, I. B. 38
Rickard, J. T. 471
Robbins, W. E. 704
Root, W. L. 441, 684
Russell, D. L. 52
Sain, M. 399
Sandberg, I. W. 875
Sandell, N. R., Jr. 860
Sannuti, P. 818, 1101
Satsangi, P. S. 896, 905, 915

Sauer, P. W. 253
Schlueter, R. A. 408, 1064
Schwartz, S. C. 517, 944
Seguin, G. E. 731
Seshadri, V. 399
Shah, T. 597
Shreve, S. 636
Sidhu, G. S. 935
Šiljak, D. D. 99, 389
Simaan, M. 359
Sinha, A. S. C. 897
Sinha, N. K. 998, 1016
Slater, G. L. 1063
Smith, J. E. 606
Snyder, D. L. 38, 219, 221, 470, 963
Sridhar, B. 974
Sripad, A. 219, 221
Suen, L. C. 423
Sullivan, R. L. 261
Sundareshan, M. K. 140
Suzuki, K. 559
Swaminathan, B. 792
Symons, F. W., Jr. 201
Talavage, J. J. 626
Tenekitzis, D. 860
Thomas, J. B. 432, 925
Thorbjornsen, A. R. 578
Toda, M. 974
Tomcik, J. D. 500
Trick, T. N. 233, 1042
Triggiani, R. 80
Ur, H. 222
Van Schuppen, J. H. 1, 470
Van Wieren, R. L. 1064
Varshney, R. K. 890
Weinert, H. L. 935
Wise, G. L. 211, 460
Wojcik, A. S. 1131
Wu, F. F. 261
Zobrist, G. W. 750
Zuhde, F. 1113

ENGINEERING LIBRARY
UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS

ENGINEERING LIBRARY
UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS

BOUND BY
HNM

UNIVERSITY OF ILLINOIS-URBANA
621.31921AL54P C001
PROCEEDINGS.$URBANA ILL
14 1976
3 0112 008033646